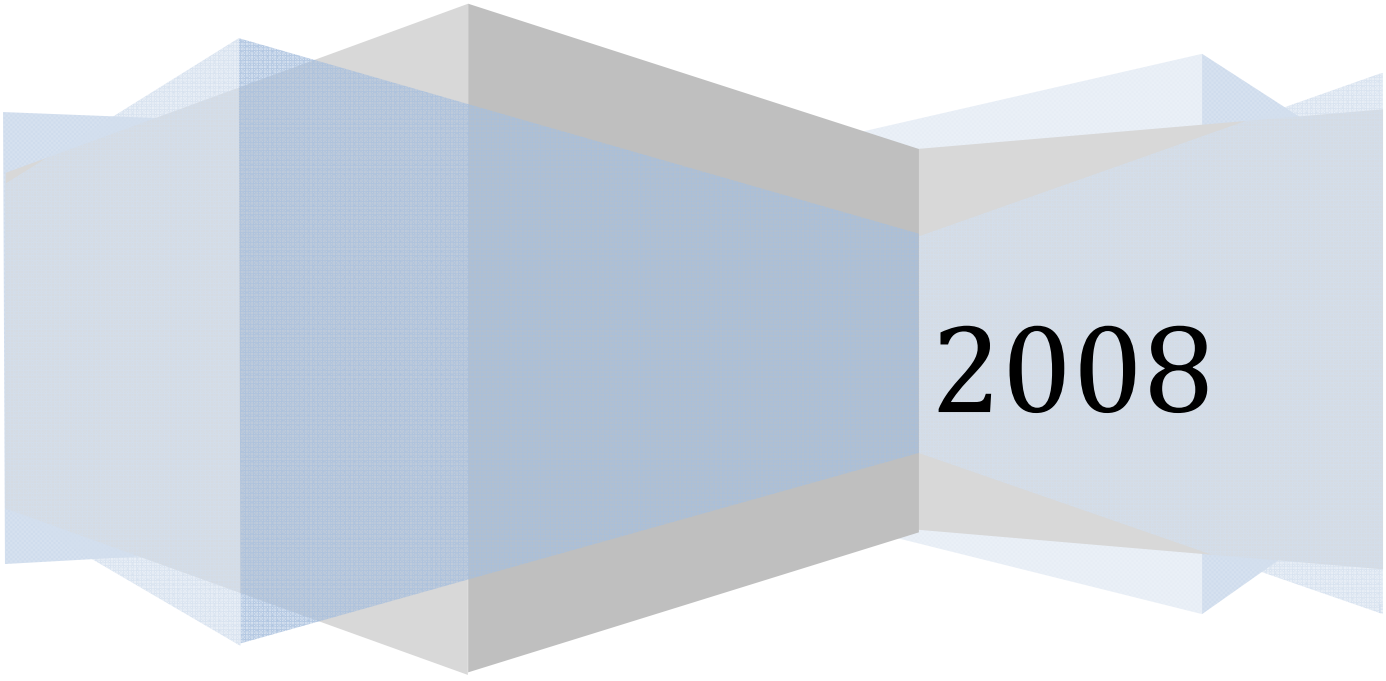
Towards Adaptive CALL

Natural Language Processing for Diagnostic Language Assessment

Iowa State University

Edited by

Carol A. Chapelle
Yoo-Ree Chung
Jing Xu



Selected Papers from the Fifth Annual Conference on Technology for Second Language Learning

ACKNOWLEGEMENTS

The papers in this volume were presented at a conference held at Iowa State University on September 21 and 22 of 2007. We are grateful for a grant from the TOEFL® Grants and Awards Program at Educational Testing Service for making this conference possible.

We thank the invited speakers for their enthusiastic participation: Robert Mislevy, Eunice E. Jang, Yong-Won Lee, Nathan Carr, Xiaoming Xi, and Mat Schulze. Although Trude Heift was unable to attend, her pioneering work in this area was cited many times.

We also thank the Departments of English and World Languages and Culture as well as the Program in Linguistics at Iowa State University for their financial support.

2008
TESL/Applied Linguistics
Iowa State University

TABLE OF CONTENTS

Introduction

TOWARDS ADAPTIVE CALL:
Natural Language Processing for Diagnostic Language Assessment

Carol A. Chapelle
Yoo-Ree Chung
Jing Xu
Iowa State University

Many advances in computer-assisted language learning (CALL) require researchers to draw upon technical knowledge about diagnostic assessment, student models, and natural language processing to design adaptive instruction. The fifth annual conference on Technology for Second Language Learning held at Iowa State University on September 21 and 22, 2007 brought together researchers and graduate students working to address issues in these areas. A day and a half of presentations, many of which are included in this volume, spanned the issues pertaining to development and evaluation of adaptive systems for second language learning. The overarching aim for the conference was to better understand the nature of adaptivity and how it can be achieved in real world applications that help language learners by assessing their language abilities and taking action based on the assessment.

The papers in this volume are divided according to four themes that they develop. The first section includes three papers discussing adaptivity. The first one is based on the paper presented by Robert Mislevy, who framed the issue of adaptivity by describing the many ways that adaptivity can be constructed in assessments. Drawing on work in *Frames of Discernment* (Shafer, 1976) and *Evidence-Centered Assessment Design* (Mislevy, Steinberg, & Almond, 2003), he proposed a taxonomy which categorizes assessments according to claim status (fixed or adaptive), observations status (fixed or adaptive), and the controlling parties of claims and observations (examiner- or examinee-controlled). The paper in this volume illustrates how the combinations of options for adaptivity appear in existing language tests and hypothetical ones that might be developed in the future. In doing so, it provides the terms and concepts needed to expand professional knowledge about adaptivity in a way that clarifies existing practice and generates new possibilities.

Joan Jamieson, Maja Grgurovic, and Tony Becker illustrate the procedure of developing and evaluating two diagnostic assessments (i.e., Readiness Check and Achievement Test) in order to support adaptivity in a commercial ESL textbook called *NorthStar*, intended to help ESL students prepare for TOEFL iBT . In this study, they investigate whether the diagnostic tests assist both the teacher and the students in preparing for the unit and in evaluating students' learning achievement at the end of the unit. The participants'

questionnaire responses suggest that pre- and post-unit diagnoses may support adaptive extension of classroom language instruction for individual students. The paper demonstrates some of the challenges of attempting to operationalize diagnosis in commercial materials.

Quan Zhang investigates the potential of applying the Cognitive Response Theory in computer-based language assessment. He suggests that computerized cognitive testing (CCT) has several advantages over computer-adaptive testing in terms of the knowledge and skills assessed, the variables taken into consideration, the task format adopted, and the scoring method used. His study of approximately 200 examinees taking a CCT using jumbled word test items reveals that some cognitive variables reflected in test-taking behavior, which can be assessed in CCT but not in traditional computer adaptive testing, distinguish examinees' levels of language proficiency. In addition, the author uses the latent factor approach to model a CCT examinee's language ability based on data collected accumulatively from a college level English test.

Central to more sophisticated adaptive systems are student models, which for language learners, need to model learners' interlanguage or state of language ability. A student model, unlike a single test score, is capable of representing a learner's detailed language knowledge based on evidence provided in their complex linguistic performance. However, if a system is to gather data to populate a student model representing language knowledge, it must be able recognize the relevant aspects of language in an examinee's responses. The second group of papers reports on the use of natural language processing in the evaluation of ESL learners' language performance.

Elena Cotos and Nick Pendar explain that many computer-assisted language tests make inferences about learners' L2 proficiency based on examinees' selected responses. They argue that the use of natural language processing (NLP) for L2 writing assessment would improve the inferences that could be drawn about learners' writing ability. They began by pointing out the advantages of constructed responses such as automatic evaluation, the provision of meaningful feedback for better learning, increased practicality and objectivity of assessment and describe what these tests look like, discussing their inherent characteristics, construct definitions, and types of test items. Finally, by reviewing several current Automated Essay Scoring Systems (AES) and approaches to natural language processing (NLP), they reveal the potentials of applying NLP techniques in automated diagnostic writing tests.

Nathan Carr discusses the relationship between decisions about automated scoring criteria and refinement of constructs in operational tests. Among three general automated scoring approaches, Carr argues for the keyword matching for comprehension test items. He illustrates how the implementation of this approach in a web-based test affected the decisions about scoring criteria and how test constructs in turn had to be altered and modified in regard to seven aspects of scoring criteria developed by Carr, Pan, and Xi (2002). The author's delineation of his ongoing development of a low-budget keyword matching program that runs in Microsoft Excel carries out the suggestion that

purposefully selected automated scoring approaches are applicable to small-scale, low- or mid-stakes diagnostic assessment of language learners' performance on target language skills and he suggests that this approach is worth exploring for well-funded, large-scale language tests.

Xiaoming Xi applies an argument-based approach for validation of the internet-based TOEFL iBT Speaking Practice test, which uses automated scoring of examinees' responses. Based on Clauser, Kane and Swanson (2002) validation framework, she builds an interpretative validity argument made up of a chain of inferential links connecting test performance to score-based interpretations and uses (modified from Kane, Crooks & Cohen, 1999 and Bachman, 2005), and then evaluates the plausibility of such an argument. This paper illustrates the process of developing such an interpretative argument and proposes the evidence needed to back up each inferential link in the argument. Meanwhile, it demonstrates the impact of automated scoring – both enhancement and potential threats – on the inferences in the complete validity argument rather than as a single consideration such as reliability.

In adaptive systems intended to help learners to develop their language ability over time, analysis of language performance is the necessary starting point, but these results in addition to other learner data need to be gathered over time, diagnostic inferences must be drawn from them, and such inferences need to be understood in terms of their meaning for language acquisition. Each of the papers in the third section addresses one aspect of this complex scenario of data gathering and interpretation.

Eunice E. Jang points out limitations of proficiency and achievement assessments in second language instruction and suggests cognitive diagnostic assessment (CDA) as an alternative. Jang asserts that CDA provides teachers with formative diagnostic information to let them refine instructional plans and also that CDA might improve students' second language learning. The summary of her research in which Jang applied CDA to iBT TOEFL preparation materials illustrates methodological, conceptual, and pedagogical challenges, which prompt suggestions for future research. For more optimizing CDA in L2 learning, Jang proposes computer-assisted CDA, arguing that it can facilitate (a) immediate reporting of diagnostic feedback, (b) authentic skill assessment, (c) utilization of various sources for diagnosis, and (d) flexible customization of diagnostic testing. A framework for computer-assisted CDA suggested by Jang depicts how the various different parties involved in L2 instruction and assessment can cooperate to enhance second language learning.

Jinhee Choo and Doe-Hyung Kim analyze data obtained from learners working on CALL to explore what variables may be considered in building an informative student model. Recognizing that a well-grounded student model may potentially contribute to second language acquisition research, Choo and Kim first quantitatively analyze learner data collected and examine possible interaction between a variety of variables such as gender, time spent on the program, proficiency level, and improvement in English. The data suggested that that learner performance may improve in relation to the amount of time

learners spend on the CALL program. The authors also analyzed the users' performance on the CALL program in relation to three different feedback types (i.e., expected, try again, and generic). Drawing upon the results of a statistical survival analysis, they suggest that the expected feedback type leads to the highest number of correct answers among the three while generic type feedback is least helpful in prompting correct answers. Choo and Kim's study exemplifies a potential use of CALL programs in establishing a student model and analyzing learner data, whose results may inform SLA research in a more sophisticated way than research on regular classroom instruction.

Mathias Schulze sketches a new approach to model student second language learning processes for individualized (adaptive) CALL systems by taking a Dynamic Systems Theory (DST) perspective to second language acquisition. In contrast to other approaches to second language acquisition, (such as the 'contrastive hypothesis', error analysis, and interlanguage analysis,) the DST approach aims to predict the next state (rather than a remote one) of a student's language learning system and provides a basis for mathematical (computational) implementation of student modeling. The Mocha project is a student second language learning model that Schulze's team intends to build by taking the DST approach. Currently, the project is at the conceptualization stage. They are experimenting with a modeling technique which borrows ideas from machine learning and are investigating the use of construction grammar for the linguistic analysis of learner text.

Melissa Baralt's study investigates the effects of computer-mediated communication (CMC) on the acquisition of L2 vocabulary items and suggests the possibility of analyzing CMC chat logs for assessment and learning purposes. In the quantitative part of the study, Baralt analyzes over 15 dyads of Spanish learners and native speakers who negotiated the meanings of 15 vocabulary items via either CMC or face-to-face interaction (FTF) mode. She found that although both CMC and FTF interaction helped beginning-level learners gain L2 vocabulary, the former outperformed the latter on improving learners' productive skills of the vocabulary items. Further, the proficiency level of the partner with whom a learner was paired did not affect the learning outcomes in either mode. In the qualitative part of the study, she invited 4 beginning-level CMC participants to review their chat logs. What she found was that learners were able to identify the places of non-understanding and errors, and reflect upon their shortcomings in their L2 ability. Thus, she suggests that saved CMC chat logs may be used for learners' self-assessment as well as for instructor's diagnosis over learners' language ability.

The papers in the final section, address important issues in language test development and validation. John Levis and Viviana Cortes point out the temptation felt by developers of diagnostic tests to select discrete aspects of language as the focus of test items. They question the utility of minimal pairs used in pronunciation textbooks, pointing out that they do not reflect syntactic contexts and frequencies of the words. In order to test their assumption that communication breakdown due to nonnative speakers' mispronunciation of a contrast found in minimal pairs is unlikely in a real-world conversation, the authors

investigated occurrences of 16 minimal pairs that involved low functional loads (i.e., /θ/ vs. /f/, /t/, and /s/) and 10 minimal pairs that involve high functional loads (i.e., /ɪ/ and /i/) from two corpora. The examinations reveal four patterns of minimal pairs regarding frequency. On the basis of the results, Levis and Cortes argue for the use of contextualized minimal pairs with high frequency words for pedagogical and assessment purposes. Finally, they suggest four hypotheses that need to be evaluated in future research.

Collectively these papers take first steps to address the interrelated issues underlying the development and evaluation of adaptive systems that include a solid measurement basis and a means of analyzing learners' language. In particular, they include some approaches and initial data on diagnostic language assessment, natural language processing for assessment, student models and complex record-keeping in language learning. Much more research remains to be done in these areas. We hope that this collection helps to focus future research on these areas that have promise for increasing the capabilities of language learning technologies in ways that have promise for helping language learners.

Part I

Adaptivity

Mislevy, R., Chapelle, C. A., Chung, Y.-R., Xu, J. (2008). Options for adaptivity in computer-assisted language learning and assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 9-24). Ames, IA: Iowa State University.

Options for Adaptivity in Computer-Assisted Language Learning and Assessment

Robert Mislevy
University of Maryland

Carol A. Chapelle
Yoo-Ree Chung
Jing Xu
Iowa State University

Levy et al. (2006) framed the issue of adaptivity by describing the many ways that adaptivity can be constructed in assessments. They proposed a taxonomy which categorizes assessments by three dimensions with potential adaptive power—claim status, observations status, and the controlling parties of claims and observations. This paper interprets these dimensions in the taxonomy in the domain of language assessment by providing language tests in the current market as specific examples. By introducing a richer concept of adaptivity, it sheds light on the development of a new generation of language assessments with the help of technology.

Most language teachers and researchers have an idea of what adaptivity means because they are acquainted with a computer-adaptive language test. What we will call a traditional computer-adaptive test determines examinees' level of ability in reading comprehension, listening comprehension or general language proficiency, for example. In such a test, examinees are presented a sufficient number of items, one at a time, for the test to make a reliable estimate of their ability with respect to the construct that the test is intended to measure. After the examinee responds to the first item on the test, the program selects subsequent items based on their responses. In general, when examinees respond correctly, the test gives them a more difficult item. When they respond incorrectly, the program selects an easier item for the next one. The traditional computer-adaptive language test is efficient at arriving at an ability estimate for examinees on a construct because each examinee spends time responding to only those items that are of an appropriate level of difficulty. Such tests are particularly welcome as placement tests when a single score is needed quickly for placement into a course and as part of a proficiency battery in which limited time is available for obtaining a score for each part of the test.

The traditional computer-adaptive language test has served well over the past decades, but as Levy, Behrens and Mislevy (2006) point out, computer technology provides a wide

range of opportunities for educators wishing to develop instruction and assessment that can help learners in a variety of ways. New options, such as adaptivity, cannot be fully understood or explored using the concepts and language of past testing practices. It is limiting to equate a powerful concept such as adaptivity with the traditional computer adaptive test, which assesses a single ability through computer selection of test items. Levy et al. explain that in order to begin to understand the new options presented by computer technology, new language and concepts are needed for conceptualizing the process of assessment. Rather than terms such as "construct," "item," and "score," for example, a richer vocabulary is needed to allow test developers to design assessments that take advantage of the capabilities of technology. Levy et al. have introduced such terms, which we use in this paper to demonstrate a range of options in computer-assisted language assessment.

REFRAMING COMPUTER-ADAPTIVE TESTING

Informally, the key questions that bear on adaptivity in any assessment are "What claims are to be made about students' knowledge or skills?" "What is the evidence that will be gathered to support these claims?" "Do either the targeted claims or the kinds of evidence change over the course of the assessment?" "If they do, who gets to decide how they change?" Levy et al. formalize these notions in terms of a space for a complex set of assessment options by introducing three characteristics of testing that can vary in ways that create different types of adaptivity and therefore are suited to different uses. First, they use the expression "observation status," the selection and presentation of items, which can be either fixed or adaptive. The common understanding of "an adaptive test" vs. "a linear test" reflects two different options for the observation status. Secondly, they point out that it is not only the items (or observations) that can be presented adaptively; the construct that the test measures can also adapt according to examinees' performance. In Levy et al.'s terms, what the test measures (the construct or multiple constructs), is referred to in terms of a claim that is to be made about the examinee. They indicate that "claim status" can be fixed as it is in the traditional computer adaptive test (CAT) but it can also be adaptive. In a measurement model, the variables for observations and the variables for students that ground claims are called the frame of discernment (Shafer, 1976), and in any kind of adaptive test the frame of discernment evolves in response to students' performance. The third dimension of adaptivity is the "locus of control," which refers to who makes the decisions about how this happens, with regard to both observations and claims. The locus of control can be with the examiner as it is in the traditional CAT in which the adaptive routine for selection of the sequence of the items as well as the claim to be made about the examinee is controlled by the examiner.

The hope for computer-assisted assessment is that the power of the computer might be applied to the need for an expanded set of test uses, but for this ideal to become reality, at the conceptual level, test developers need to be able to see ways of moving beyond the traditional CAT that is useful for placement and proficiency testing. This paper aims to work toward this goal by illustrating some of the options for adaptivity of existing

language tests thereby expanding the potential for designing future language tests to suit their specific purposes. The actual and hypothetical tests that we discuss are displayed in Table 1, which shows the position of each in a three dimensional space delineated by observation status across the top, claim status along the vertical, and locus of control on the third dimension. The third dimension is shown in this two-dimensional figure by dividing each of the positions of the vertical and horizontal. We will discuss each of these tests in turn, describing the ways that their observation, and claim status each make them either fixed or adaptive, and discussing the differences in examiner vs. examinee control for each.

CLAIM STATUS

The claim for a test refers to the statement that is to be made about the examinee on the basis of observations of his or her performance on the test. A claim might be that an examinee has strong reading comprehension ability or that the examinee is able to write an effective essay using the conventions of standard written English. In each of these.

Table 1. Examples of language assessments with a variety of types of adaptivity

Observation Status		Fixed		Adaptive	
Claim Status		Examiner Controlled	Examinee Controlled	Examiner Controlled	Examinee Controlled
Fixed	Examiner Controlled	(1) CET-4 WT; RCAA (Jamieson et al., this volume)	(2) Transparent Language Test	(3) ACT EPT	(4) Hypothetical grammar test with item-level feedback
	Examinee Controlled	(5)	(6) nonsensical	(7)	(8)
Adaptive	Examiner Controlled	(9) SOPI	(10) nonsensical	(11) OPI	(12) Hypothetical CGT based on CCT (Zhang, this volume)
	Examinee Controlled	(13) FETN; S-TOPIK	(14) nonsensical	(15) DIALANG	(16) Hypothetical online language test site

cases, the claim refers to one ability or multiple abilities of the examinee, but it would not be difficult to imagine a test in which different claims about language ability are made about examinees depending on their performance. For example, what if an examinee performs poorly on the part of the test requiring recognition of correct grammatical forms and is therefore not required to spend time completing the essay. In this case, the claim for some examinees would be limited to grammar, whereas for others, the claim would be about writing ability including a claim about grammar knowledge. In other words, adaptive claims are possible depending on the examinee's performance. The tests in Cells 1-4 in Table 1 contrast with the rest of the examples in that the former are intended for making a fixed claim or claims about the examinees, like the traditional CAT does, whereas the latter may result in adaptive claims.

Tests with Fixed Claims

Table 1 shows examples of tests with fixed claims in Cells 1 through 4. In Cell 1, the College English Test Band 4 Written Test (CET-4 WT)[i] is a high-stakes test intended for certifying college students' general English proficiency in China. The test makes a single claim that is fixed based on the examiner's choice of intended test inferences. The test consists of six sections—Writing, Skimming and Scanning, Listening, Cloze, Reading in depth, and Translation—and they are intended for assessing four language aspects selected by the examiner: writing, reading (skimming and scanning), listening, and integrated language ability which is comprised of intensive reading, Chinese-English translation, and vocabulary and structure. Based on an examinee's performance in these four language aspects, a total scaled score and a percentile rank—as compared to other examinees—are reported as the indicator of his/her general English proficiency. The observables of the CET-4 WT are also fixed and controlled by the examiner. As a paper-and-pencil test, the CET-4 WT presents the same test items in a predetermined order to all examinees. The task types used in the test include essay writing, true and false, cloze, multiple-choice, and fill-in-the-blank. This test contains no adaptive elements so each examinee is given the same amount of time to complete each section.

Also in Cell 1, the Readiness Check and Achievement Assessment (RCAA) developed by Jamieson et al. (this volume) are low-stakes tests which have fixed claims as well as fixed sets of observations controlled by the examiner. The RCAA tests are intended to help teachers and learners at university-level intensive English programs to understand areas of language knowledge that learners need to work on. The Readiness Check test and the Achievement test. The two tests are used for different pedagogical purposes. The first is used to check students' readiness for in-class instruction and activities while the second is to assess students' learning outcomes after a class. The constructs of these two tests are fixed and are determined by the examiners based on their analysis of what students have studied, what they are going to study in their language classes, and the difficult language aspects that previous students reported. The observations of the tests are also fixed and selected by the examiners as students enrolled in the language learning program are always presented the same test items. However, based on the test results, students are provided with individualized remedial materials. Although RCAA and CET4-WT fall in

the same cell, they are different in terms of the number of claims the test can make. In CET-4 WT, only a single claim is made about examinees' general language proficiency. By contrast, RCAA makes multiple claims in both vocabulary and grammar knowledge (Jamieson et al., this volume). From this example, we can see tests with fixed claims can have more than one claim and that tests in a single cell can be for high stakes or low stakes.

Cell 2 contains a low-stakes counterpart to the CET-4. Delivered on the Web, the Transparent Language English Proficiency Test (TLEPT)[ii] for native Spanish speakers allows examinees to choose the area of their general English proficiency to be tested. In this sense the observations are examinee-controlled. Once that initial choice is made, both the claims and observations are fixed and examiner controlled. The test consisting of four sections is intended to make claims about three language aspects: grammar knowledge, vocabulary knowledge, and reading ability. Two grammar sections of the test assess an examinee's ability to manipulate sentence elements (verbs, adjectives, prepositions, conventions, modifiers, and function words) as well as to recognize erroneous sentence elements. A vocabulary section, on the other hand, assesses an examinee's ability to select and use appropriate words in a given context. Finally, the reading section assesses an examinee's referring, inferring, and summarizing ability in reading. The score of each section is reported individually in terms of the percentage of items answered correctly. Though the test items of the four sections are predetermined by the examiner, the examinee has the freedom to choose the way to proceed through the test and to select the order in which parts of the test and individual multiple choice items are completed. The examinee is also free to spend as much or as little time on the test as he or she wishes. Even though this test is not adaptive in terms of claims or observations, it provides for some elements of examinee choice in how the test is completed.

As noted in the introduction, the most familiar adaptive tests lie in Cell 3. The kind of adaptivity that characterizes this cell is evident in the ACT ESL Placement Test (ACT-EPT).[iii] The ACT-EPT is a medium-stakes test intended for placing postsecondary students into appropriate ESL courses in the United States. It has fixed, examiner controlled claims and adaptive, examiner controlled observations. The test consists of three modules and each is intended to make a claim about one aspect of language ability selected by the examiner: grammar/usage, reading, and listening. The grammar/usage module assesses an examinee's ability to recognize and manipulate sentence elements (verbs, subjects and objects, modifiers, function words, conventions, and word formation), and sentence structure and syntax. The reading module assesses an examinee's referring and reasoning ability in reading and the listening module an examinee's ability to understand explicitly and implicitly stated information in speech. The score on each module is reported in terms of five levels ranging from near-beginner to near-native speaker. Detailed proficiency descriptors are also provided to define the things that a typical student at each proficiency level can do. The observations on the ACT-EPT are adaptive based on each learner's language ability and are controlled by the adaptive procedures in the test which were designed by the examiner. Like the traditional CAT,

the ACT-EPT presents multiple-choice test items to an examinee based on his or her previous responses and thus routes the examinee to the appropriate levels of test items until a sufficient number of items has been given at the appropriate level.

An examinee-controlled counterpart to the traditional CAT is adaptive with examinee controlled observations. In other words, it is the examinee who selects which items to complete on the test in order to obtain a score. Although we could not find an existing language test as an example for Cell 4, we can imagine a hypothetical Cell 4 grammar test, which might be useful for instruction and learning. Such a grammar test would provide test takers with feedback on their performance on each item as illustrated by the program developed by Choo and Kim (this volume). In addition to providing feedback for test takers, let us say this grammar test allows test takers to use the feedback they receive on each item to help them in selecting the difficulty level of the following item. When the test is finished, a total test score is computed on the basis of the difficulty level of selected test items that were answered correctly. The claim of such a grammar test is fixed by the examiner because it is intended to make a single claim about examinees' grammar ability, producing a single test score. At the same time, such a test inference is chosen by the examiner while the observations are adaptively determined by the examinee. Accordingly, although the test item pool is pre-determined by the examiner, individual examinees may encounter different items during test taking depending on their own observations and judgments. In this scenario, the hypothetical grammar test is a low-stakes assessment, which may be useful for instruction.

All of these tests with fixed claims have claims that are defined by the examiner, who developed the test to measure a construct or constructs such as reading comprehension. But as we saw, a single claim about reading comprehension or vocabulary knowledge, for example, can be arrived at through a fixed set of observations that is invariant across examinees regardless of their performance (Cells 1 and 2), or it can be made on the basis of observations selected adaptively (Cells 3 and 4). Even when claims and observations are fixed, some element of learner control can come into play in settings where examinees have choices about what to be tested on and the order they wish to complete the items (Cell 2). Adaptivity can be controlled by a set algorithm designed by the examiner to select items on the basis of prior performance by examinees (Cell 3), or it can be controlled by the examinees themselves (Cell 4). In the examples from Cells 1 through 4, we saw the most traditional linear (Cell 1) and adaptive (Cell 3) tests, but we also saw tests that expand the test use into instruction and learning by providing choice and immediate feedback about performance to the learner.

Tests with Adaptive Claims

In many tests, examinees' performance is interpreted to make claims about different constructs at different levels. For example, beginners may demonstrate performance that provides a basis for claims about vocabulary and pronunciation alone, whereas at an advanced level, performance would allow for claims about these aspects of language in addition to rhetorical knowledge. Tests yielding claims about more than a single aspect of

language have the potential for adjusting the constructs tested during the testing process. In a language test, the student model variables concern the aspects of language ability to be assessed while the observable variables concern learner behaviors which provide evidence for levels of proficiency in these language aspects. In the example given above, the frame of discernment reflects choices of the test constructs among vocabulary knowledge, pronunciation, and rhetorical knowledge and such choices are made based on learner performance (observations). Now both the student model variables and the observable variables can play a role in adjusting claims and yielding interpretations of learner performance in an intertwined manner as an adaptive-claim language assessment proceeds. At this point, it may be appropriate to draw clear relationships between the number of claims and the adaptivity of claim status before moving on to discuss examples of multiple-claim assessments.

It is witnessed earlier that the relationship between the dimensionality (or multiplicity) of claims and the adaptivity of claim status is not of a one-to-one relationship. Fixed claims may comprise one or more claims, but this is not true for adaptive claims. In this regard, Levy et al. (2006) point out three general properties of assessment as follows:

(a) Adaptivity of claims in assessments entails multiple claims;
(b) Univariate-claim (i.e., single-claim) assessments are inevitably fixed; and
(c) Multiple-claim assessments can be either fixed or adaptive (p. 6).

These relationships between multiplicity and adaptivity of claims in assessments are shown in Figure 1. Note that single-claim assessments are always fixed, but not vice versa. As a single-claim language assessment focuses on only one aspect of language ability, adaptivity in claim status, which is by and large constructed through multiple observations, does not play a role in such univariate assessments. On the other hand, multiple-claim assessments are either fixed or adaptive. As seen below, a fixed multiple-claim assessment can involve either fixed or adaptive observations but always addresses the same set of claims, adaptive multiple-claim assessment may involve multiple observations as 'the hypotheses of interest that are investigated may change as new information (from observation) is brought to bear' (Levy et al., 2006, p. 5).

Examiner-controlled adaptive claims

Cells 9 through 16 of Table 1 provide examples of tests with adaptive claims. From these examples, we see the choices of the test constructs can be either made by the examiner or the examinees. Three examples illustrate tests whose adaptive claims are controlled by the examiner. In Cell 9, the Simulated Oral Proficiency Interview (SOPI),[iv] a speaking proficiency test that yields a single score which can be used for a variety of purposes, has adaptive, examiner-controlled claims and fixed, examiner-controlled observations. The SOPI consists of four parts: warm-up, level checks, probes, and wind-down. After responding to warm-up questions, an examinee's proficiency level is evaluated via tasks designed for level-checking and observation. Trained raters make claims about the examinee's oral proficiency by listening to his or her recorded responses to given

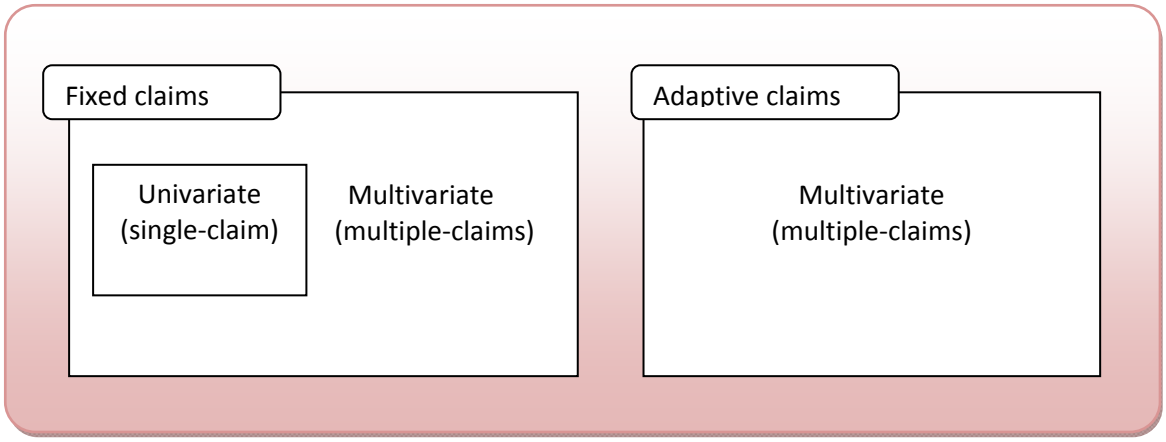


Figure 1. Relationships between the dimensionality of claims and the adaptivity of claim status

prompts in regard to different language functions described in the ACTFL Guidelines for speaking[v], which classifies proficiency levels into Novice, Intermediate, Advanced, and Superior. Scores for the SOPI range from Novice-mid to Superior and the claims associated with each of these levels differ. For example, the Novice level makes claims about examinees' vocabulary, oral fluency and the complexity of speech while the Superior level makes claims about the accuracy, pragmatic competence, and interactive strategies. Whereas the claims vary by level, the observables of the SOPI are fixed and controlled by the examiner. Developed as a semi-direct equivalent test to the face-to-face Oral Proficiency Interview (OPI), the SOPI delivers test items from a recorded tape and a test booklet. Since all test items are identical across the test takers, the SOPI can be administered to a group of examinees in a language lab setting simultaneously.

In Cell 11, the ACTFL Oral Proficiency Interview (OPI)[vi] is a face-to-face or telephone-mediated oral test. Like the SOPI it has adaptive, examiner-controlled claims, but because if is given as a one-on-one interview, the observations can be chosen adaptively by the examiner, as well. Claims about an examinee's oral proficiency are made adaptively during the test administration by a trained interviewer, who also rates the examinee's performance with another trained rater. The interviewer starts with items targeting a certain proficiency level and adjusts the difficulty level of test items as the test goes on, by changing topics and language functions, on the basis of the interviewer's assessment of the examinee's performance on test items. Test items may be presented in the form of natural communication or role-plays. If the examinee does not feel comfortable about a certain topic, he or she can request a topic change. The interviewer controls the test-taking time on the basis of his or her perception about the examinee's proficiency level. Like the Simulated Oral Proficiency Interview in Cell 9, the examinee's performance may be rated using either the ACTFL proficiency guidelines or the 11-point ILR scale[vii], ranging from Novice (0+) to Superior (3 and above).

For Cell 12, one can imagine a grammar test with adaptive claims determined by the

examiner and observations selected by the examinee. Such a hypothetical cognitive grammar test (CGT) is an extension of the computerized cognitive test (CCT) illustrated by Zhang (this volume) in which examinees select cues that help them in responding to test questions. In the hypothetical CGT, the examiner may adjust the inferential targets—though still with a primary focus on grammar knowledge—depending on the cues examinees choose to solve the language problems. For example, if an examinee does much better in jumbled-word test items when they get help from metalinguistic cues—such cues identify the stem sentence (subject, verb, and object) in a complicated sentence structure – than when they skip such help options, the examiner may decide to make claims about the examinee's metalinguistic competence in addition to his or her grammar knowledge. Such claims may be that the examinee displays high competency of metalinguistic knowledge but low grammar knowledge. Likewise, if the examinee has to rely on cues of key word definitions to assemble jumbled words into meaningful sentences, the examiner will then include vocabulary knowledge into the test claims. In this case, the test claims may be that the examinee shows high level of grammar knowledge but low level of vocabulary knowledge. The observations of the CGT are also adaptive but are subject to examinees' control. While the cues are provided by the examiner, examinees may choose either types of cues or decide not to use any cues during test taking.

Examinee-controlled claims

The adaptive claims of a language test might also be selected by examinees, who choose the claims that they wish to be able to make about their language ability. Examples of such tests appear in Cells 13, 15, and 16. In Cell 13, the Free-English-Test.Net (FETN)[viii] is a low-stakes English test website for English as second language (ESL) learners to assess their English proficiency in various specific aspects of language ability at three different difficulty levels. It has adaptive, examinee controlled claims and fixed, examiner controlled observations. The claims made by FETN are adaptive based on the examinee's choices of five major sections, including three levels of grammar, synonyms, business English, usage, and idiomatic expressions. Under each section is a large number of sub-tests, each assessing one specific language aspect and each categorized into one of the three difficulty levels: elementary, intermediate, and advanced. Upon entrance to the test website, an examinee has the freedom to choose any sub-test under a certain section and at a certain difficulty level to receive claims about a specific language area and level. The FETN test website in its current form is limited to item-level feedback. However, if such a test could provide summative evaluation for examinees on each specific aspect of language, the score would better serve as an overall claim. As an internet-based test, FETN presents the same test items in each sub-test to all examinees and consistently uses the multiple-choice question format.

A high stakes example of a test in Cell 13, the Standard Test of Proficiency in Korean (S-TOPIK)[ix] assesses Korean as a foreign language learners' general Korean proficiency primarily for admission and hiring purposes. It has adaptive, examinee controlled claims and fixed, examiner controlled observations. The S-TOPIK makes claims that are

adaptive based on examinees' choices among three proficiency levels, i.e., beginner, intermediate, and advanced, at the beginning of the test. Scores are then contingent upon examinees' initial choices of proficiency levels. Each level of the S-TOPIK consists of four sections: vocabulary and grammar, writing, listening, and reading. The score of each level is reported in terms of a standardized total score, which corresponds to a lower or upper band of the level. The Korean proficiency of an examinee thus can be interpreted by looking up the descriptions of the band in which his or her score falls. Once one of the three levels has been chosen by the examinee, the examiner controls a fixed procedure of obtaining observations. The test presents the same task items in a predetermined order to examinees who select the same level of the test. Test items in all sections except writing are given in a multiple-choice format, each presenting 30 questions. Different limits regarding the length of the composition are posed to examinees based on their target proficiency level. An examinee is required to complete the writing section in an hour and each of the other sections in 30 minutes. Thus, the entire test lasts for three hours.

Compared with S-TOPIK, DIALANG's online diagnostic language testing system (DIALANG)[x] in Cell 15 has the same claim status but different observation status. This low-stakes test intended for informing language learners about their proficiency levels as well as providing tips for language learning has adaptive, examinee controlled claims and adaptive, examiner controlled observations. Similar to S-TOPIK, DIALANG makes claims that are adaptive based on examinees' choices of languages and language aspects to be assessed. The online testing system offers assessments of fourteen languages in five language aspects, including listening, writing, reading, structures, and vocabulary but the examinees make decisions on what will be assessed when entering the test. DIALANG reports examinees' ability in each language aspect in six levels ranging from beginner (A1) to very advanced (C2) based on the Common European Framework. In addition, the test provides detailed score descriptions and suggestions for each level of learners. In contrast to S-TOPIK, the examiner controls the observations which are adaptive to examinee's responses during the assessment. Based on examinees' performance in a placement in which they are asked to distinguish between real words and pseudo-words and their responses to an optional self-report of language ability, the examiner routes the examinees to the appropriate levels of test items. Such routing activity always continues in the testing process. Thus, examinees of different levels of language proficiency encounter different test items.The task types used in the test include multiple-choice, gap filling, sentence completion, sentence insertion, error recognition, and word formation. The DILANG test is not timed, and examinees are allowed to spend as much time as they want to on the test.

An example for Cell 16 would be a hypothetical online site[xi] a language test system whose claims and observations are adaptive and controlled by the examinees. Imagine a website which has a variety of language tests in its database. A single test is categorized in regard to the target language abilities and difficulty level. To take a test, examinees visiting the website would type in the search box the name of a language skill that they want to be assessed (say, vocabulary). The search engine would retrieve all the tests that

make claims about vocabulary ability from its database and list them on the screen. The retrieved tests may be sorted by the difficulty level, sub-constructs (such as parts of speech, collocations, and idioms), or topics (such as hospital, cooking, school, travel, culture, etc.). Examinees then read through the list of vocabulary tests and select what they are interested in. They may choose to take more than one vocabulary test in various orders. Examinees may also quit the test in mid course and switch to another test in the list. In this scenario, claims are controlled by examinees in an adaptive manner perhaps informed by feedback they receive. As there is no restriction in selecting the target construct in addition to difficulty levels, examinees may enjoy freedom in searching for the language tests they wish to take on the website.

OBSERVATION STATUS

In the traditional CAT, the observation status is the dimension that is adaptive, and this dimension is controlled by the examiner through the selection of items based on an algorithm that considers the examinee's prior performance. We saw from the examples in cells beyond Cell 4 that adaptivity does not have to refer to adaptive observations, but can also refer to adaptive claims. The SOPI in Cell 9 and the FETN and S-TOPIK in Cell 13, for example, make adaptive claims but have fixed observations. The claims made by these three tests about examinees are adaptive and controlled by either the examiner's or examinees' choices of specific language abilities to be assessed. The adaptive claims made by SOPI are subject to the examiner's decision. Through a level-checking process of the test, the examiner places examinees into appropriate proficiency levels, each of which makes claims about different aspects of language ability. In contrast, the adaptivity of claims in FETN and S-TOPIK are controlled by examinees. In these two tests, examinees are entitled to decide the proficiency level or aspects of language ability to be assessed and, accordingly, the examiner makes corresponding claims. With these adaptive claims, the three tests have fixed observables which are controlled by the examiner. Once the scope of claims (or a "claim space" in Levy et al.'s terms) is determined, the examiner then presents a set of prearranged test items to examinees regardless of their test performance.

Although tests with adaptive claims and fixed observations exist for language assessment, Levy et al. point out that fixed observations are in many cases insufficient for assessment with adaptive claims because the predetermined fixed observables maybe optimal for assessing one part of the claim space yet inadequate for the other parts. In other words, an assessment having fixed observations may have limited flexibility to adjust its focus in the claim space. Such a drawback can easily be detected in the SOPI in Cell 9. Though the test has adaptive claims, its adjustment of focus in the claim space can take place only once and only at the beginning of the assessment before many observations have been gathered. Such a limiting adaptivity of test claims is largely attributed to the fixed test format which prohibits the assessment from calling for optimal observables to make specific claims and thus restrains the assessment from moving around the claim space freely in the testing process.

In contrast, language assessments having adaptive claims as well as an adaptive observation status can adjust the claims multiple times while flexible observables about the examinees are collected. Taking the hypothetical extension of CCT in Cell 12 as an example, the test is able to shift its focus in the claim space onto any of the three aspects of the examinee's ability, including grammar knowledge, metalinguistic competence, and vocabulary knowledge, based on the examinee's responses and uses of cues. Thus, the adaptivity of claims in language assessment can typically be better operationalized when the observables are adaptive as well.

LOCUS OF CONTROL

The locus of control for adaptivity in the traditional computer-adaptive language test is the examiner, who sets the algorithm for item selection. Even though the algorithm includes information about the examinee's performance, the examinee does not have any explicit choice in the selection of observations. However, from Table 1, we saw that the locus of control is a third dimension that intersects not only the observation status but the claim status as well. Thus, the locus of control with regard to claims is another variable that distinguishes language assessment. For example, both FETN in Cell 13 and SOPI in Cell 9 have adaptive claims and fixed, examiner-controlled observations. The difference between the two tests lies in that the former allows examinees to select the language aspects to be assessed—examinee-controlled claims—but the latter reserves such a job for the examiner—examiner-controlled claims. Similarly, the choice of observations to be gathered during the assessment can be made either by the examiner or examinees. For example, although a traditional computer-adaptive test such as the ACT EPT in Cell 3 and the hypothetical grammar test in Cell 4 both have fixed, examiner-controlled claims and adaptive observations, the former empowers the computer (examiner) to select test items for examinees based on their performance—examiner-controlled observations—while the latter endows the examinees the freedom to decide the difficulty level of upcoming test items based on the feedback generated by the examiner—examinee-controlled observations.

Although the dimensions of claims, observations, and locus of control can theoretically intersect with each other, some combinations of these three dimensions have not been explored (or are not applicable) yet in the context of language assessment. For example, we did not find any existing language test for the Cells 5-8 which have fixed, examinee-controlled claims. According to Levy et al., such types of assessments are nonsensical because examinees do not have any control over the claims when the claim space (intended construct) is already fixed. Further research or further reflection may reveal assessments that do fit into these cells nevertheless. For the same reason, we did not find any examples for Cells 10 and 14 in which observations are fixed yet controlled by examinees. However, the Transparent Language Test in Cell 2 is an exception. In such a test, although all examinees encounter the same form of test items selected by the examiner, they are not required to follow the given testing procedure or complete all test items. Thus, different examinees may provide different observations (complete different

number of items and in a different order) to the assessment. In this case, the examiner and examinees share the right to make choices of observations. We categorized this test as a type with fixed, examinee-controlled observations because examinees have certain freedom on choosing observables. From this example, we see that cooperation between examiner and examinees in selecting test items is also possible, particularly in this low stakes test.

Table 1 reveals that the locus of control is a dimension that can be related to the stakes of language assessment. Specifically, tests under the examiner-controlled categories are of higher-stakes than those under examinee-controlled categories. Such distinctions can be easily detected by comparing pairs of language tests in two neighboring cells, such as the CET-4 vs. Transparent Language Test, SOPI vs. FETN, and OPI vs. DIALANG. The two tests in these pairs only differ in the locus of control for one dimension but they have very different stakes. In these cells, the tests more controlled by the examiner, such as the CET-4, SOPI, and OPI are widely used high-stakes test while those controlled by the examinees are low-stakes free online assessments.

Another such pair of examples is the TOEFL Internet-based Speaking test and its counterpart, the Online Practice Speaking Test (See Xi in this volume). The two tests, although having different test constructs and purposes, share the same test format. While doing the practice test to prepare for the real test, examinees may choose an untimed testing mode in which they are allowed to read and listen to testing prompts multiple times, prepare for their speech for as long as they want to, or quit and continue the test at any time. These options of examinee control are not available in the real test. Both tests have fixed observations as they present the same test items across examinees. However, the observations are controlled by the examiner in the real test and by the examinees in the practice test. Compared with those taking the real test, the examinees doing the practice test enjoy the freedom of proceeding through the test at their own pace. From the examples above, we saw that language assessments having examinee-controlled claims or observations are often used for self-assessment or practice purposes. Thus, one of the future directions for computer-adaptive language assessment will be to develop examinee-controlled tests to prepare learners for high-stakes test purposes. The Hypothetical Online Language Site in Cell 16 is a model for this type of test. In such a test site, examinees are not only permitted to choose the construct to be measured but select test items (observables) based on interest as well.

However, if the responsibility of deciding the constructs of a language test is completely put in examinees' hands, the examinees accustomed to examiner-controlled tests may not feel empowered but bewildered instead because of their lack of knowledge about their own language ability. Suppose that a learner of Chinese hopes to see claims about his Chinese proficiency. Given the privilege to select the language skills to be assessed—possibly including the abilities to spell *pinyin* for Chinese characters, to recognize the meaning of Chinese characters, to pronounce tones correctly for given words, to order the strikes for writing a Chinese character, etc.—he will probably feel at a loss. Thus, in many cases, it might be a good idea for the examiner and examinees to share the job of

selecting the language abilities to be assessed. For example, examinees may receive feedback and suggestions from the examiner about what should be tested based on their performance and then make decisions accordingly. In other words, the examiner will guide the examinees through the test yet the examinees will still have the control over the target inferences of the test.

Tests with adaptive, examinee-controlled claims will very likely need the examiner to provide examinees with individualized feedback based on their performance. For the computer to offer such feedback, it has to be equipped with the ability to diagnose responses, such as examinees' constructed responses (see further discussion in Cotos and Pendar, this volume). In addition, the computer examiner must rely on student models to make decisions or provide suggestions on the language aspects to be assessed. Such models define the important variables related to the language ability of the examiner or the examinees' interest (Mislevy, Steinburg, Almond, & Lucas, 2006). In addition, student models are the key to providing useful feedback to examinees as they are informed by the analysis of learner text (see further discussion in Schulze, this volume).

CONCLUSION

The three dimensions of claim status, observation status, and locus of control elaborate the meaning of adaptivity beyond the one dimensional concept of a test capable of making a fixed examiner-controlled claim or claims based on an examiner-controlled set of observations. The three dimensions provide a space to recognize the adaptivity inherent in existing assessments that are used across purposes. In the examples, we saw not only the traditional linear and computer-adaptive test, but also the range of options provided by a rich concept of adaptivity. Such a rich concept allows test developers to consider a range of potential types of adaptivity for assessments that are used in proficiency and placement, as well as in achievement and diagnosis. Besides, it provides language for analyzing the ways that a test may be adaptive in some ways but not others, to meet the purpose of the test, and provides support for determining the measurement models and adaptation algorithms that best suit these purposes. The richer concept of adaptivity thus allows for assessments that are useful for the benefit of educators who wish to place and evaluate examinees as well as for learners who wish to better understand what they know and what they need to work. With such a range of assessments defined by these options for adaptivity, test developers can better consider the ways in which technology can be used to develop a new generation of language assessments.

REFERENCES

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, and assessment*. Cambridge: Cambridge University Press.

Levy, R., Behrens, J. T., & Mislevy, R. J. (2006). Variations in adaptive testing and their online leverage points. In D. D. Williams, S. L. Howell, & M. Hricko (Eds.), *Online assessment, measurement, and evaluation* (pp. 180-202). Hershey, PA: Information Science Publishing.

Mislevy, R. J., Steinburg, L. S., Almond, R. G. & Lucas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, N. J.: Lawrence Erlbaum Associates.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

---

[i] Though a well-recognized English test in China, CET-4 WT does not have its official website. Even so, many test preparation websites in China provide detailed information about the test. A sample test provided by *QQ CET-4 Test Preparation Center* can be found at http://edu.qq.com/a/20071204/ 000124.htm.

[ii] For more information about the *Transparent Language* English Proficiency Test, please visit its official website at http://www.transparent.com/tlquiz/proftest/esl/tlesltest.htm.

[iii] For more information about the ACT ESL Placement Test, please visit its official website at http://www.act.org/esl/overview.html.

[iv] Developed as an equivalent of the face-to-face Oral Proficiency Interview (OPI), SOPI is a semi-direct speaking test, administered in places where a limited number of trained raters are available for the test. More information can be found in the following documents on the website of the Center for Applied Linguistics: *Simulated Oral Proficiency Interviews: Recent Developments (*http://www.cal.org/resources/digest/0014simulated.html) and *Testing/Assessment: Simulated Oral Proficiency Interviews* (http://www.cal.org/topics/ta/sopi.html).

[v] The ACTFL Guidelines for speaking can be found at www.sil.org/lingualinks/languagelearning/OtherResources/ACTFLProficiencyGuidelines/contents.htm

[vi] Although it is a communicative oral proficiency test administered face-to-face, the OPI test has many drawbacks such as inefficient test administrations, requirement of a great number of trained interviewers, and lower reliability. To solve such problems, different types of semi-direct oral proficiency tests equivalent to the OPI have been developed, one of which is the Simulated Oral Proficiency Interview. Currently, the Computer-mediated Oral Proficiency Interview is being developed with the feature of examiner-controlled adaptive observations. For more information about the OPI development and its scoring rubrics, please visit the following two websites: Defense Language Institute (http://dlielc.org/testing/opi_test.html) and Center for Applied Linguistics (http://www.cal.org/resources/digest/oralprof.html).

[vii] The ILR scale and oral proficiency descriptions can be found at http://www.govtilr.org/ILRscale2.htm.

[viii] For more information about the FETN test website, please visit its homepage at http://www.english-test.net/.

[ix] For more information about the Standard Test of Proficiency in Korean, please visit its official website at http://www.topik.or.kr/guide/topik_en_01_d.html.

[x] For more information about DIALANG, please visit its official website at http://www.dialang.org/intro.htm.

[xi] This hypothetical test is created based on Levy et al.'s explanation of a non-language test: "Consider a simple case where a user's query results in a list of documents, possibly structured by

some criterion such as perceived relevance. The user then selects some of the documents from the list for further consideration. A great deal of observable information can be collected from such a process. Which documents were viewed? In what order? How much time did the user spend reading each? These only scratch the surface of what data could possibly be collected. In these systems, the user is in control of the claim space, via the query, and the observables, via the actions taken with respect to the produced list of documents" (p. 29).

Jamieson, J., Grgurovic, M., & Becker, T. (2008). Using diagnostic information to adapt traditional textbook-based instruction. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 25-39). Ames, IA: Iowa State University.

Using Diagnostic Information to Adapt Traditional Textbook-Based Instruction

Joan Jamieson
Northern Arizona University

Maja Grgurovic
Iowa State University

Tony Becker
Northern Arizona University

Although *diagnostic assessment* has traditionally been defined as a highly specialized procedure for addressing persistent learning problems, applied linguists including materials developers recently have associated this term with corrective procedures in formative assessment. This latter sense was used by developers of the *NorthStar* textbook series in which on-line assessments are being used in the process of adaptive instruction for English language learners. In this article, the design, development, and initial beta testing of a prototype Readiness Check and Achievement Test are described. In summer, 2007, data were collected on test performance and attitudes through scores, a questionnaire, and interviews. Analyses were conducted to examine student performance as well as the degree to which the teacher and the students found the Readiness Check and the Achievement Test helpful. Overall, these materials have apparent advantages for both students and teachers, although studies conducted with a larger and more representative sample are needed before claims regarding the benefits of the diagnostic use of these assessments to support adaptivity in *NorthStar* can be supported. This article describes a real-world example of a small scale, modest step forward for diagnostic language assessment used in instruction.

What is diagnostic assessment? It seems that the meaning of this term is interpreted differently by assessment specialists as shown in Table 1. For some, diagnostic assessment is seen as a use of formative assessment. For others, diagnostic assessment is seen as a highly specialized procedure. Some applied linguists refer to the identification of learners' strengths or weaknesses in general (Alderson & Hakuta, 2005) or in regard to students' learning in a particular curriculum (Bachman & Palmer, 1996; Chapelle & Douglas, 2006). In these examples we see diagnostic testing as a form of formative assessment. Only the non-applied linguists (Linn & Miller, 2005) and the authors of dictionaries in applied linguistics (Davies et al., 1999; Richards & Schmidt, 2002) describe diagnostic assessment as fundamentally different from formative assessment:

> To use a medical analogy, formative assessment provides first-aid treatment for simple learning problems, and diagnostic assessment searches for the underlying causes of those problems that do not respond to first-aid treatment. Thus diagnostic assessment is much more comprehensive and detailed. It involves the use of specially prepared diagnostic tests as well as various observational techniques. (Linn & Miller, 2005, p. 36)

Yet even these authors explain that formative assessments are often used for diagnostic purposes. There does appear to be general consensus that whether formative assessments or diagnostic assessments are administered, two important uses are to provide helpful information to the learner regarding his or her lack of understanding of important information (e.g., Alderson & Huhta, 2005; Chapelle & Douglas, 2006) and to inform syllabus design or course placement (e.g., Bachman & Palmer, 1996; Richards & Schmidt, 2002). From both perspectives, the concern is to be able to adapt instruction to make it appropriate for learners.

In this article, diagnosis includes two meanings. First, it refers to identification of learning weaknesses through formative assessment for the purpose of adapting the learning path to include individual remediation for some students. Additionally, it refers to identification of prerequisite skills through a readiness pretest, again with the purpose of adapting learning paths to include individual remediation for some students (Linn and Miller, 2005).

NORTHSTAR ADAPTIVITY PROJECT

NorthStar (3rd edition) is a five level textbook series for English language learners who are young adults/adults. Each of the five levels has two integrated skills textbooks—one for listening and speaking and the other one for reading and writing. Each text has ten units; each unit is divided into three sections: 1) Focus on the Topic; 2) Focus on Listening/Reading; 3) Focus on Speaking/Writing. The units are thematically based on high-interest, and somewhat controversial, topics.

The *NorthStar* Adaptivity Project was intended to support individualized instruction of English language learners. Classroom instruction, using the textbook series *NorthStar*, was planned to be adapted in two ways: 1) through on-line materials that could be individually assigned to students by their teachers, and 2) through assessment—specifically with its related feedback and remediation on the basis of learners' performance on a Readiness Check (i.e., the readiness pretest) and an Achievement Test (i.e., formative assessment). The on-line materials, assessments and their remediation were designed to be delivered on-line using an assessment management and learning platform named Pegasus. As implemented in this project, the on-line site was called MyEnglishLab and was referred to as MEL.

By identifying learners' weaknesses through assessment and by providing those who needed additional help with extra explanations and practice, it was hoped that their overall understanding and performance of the content covered in each textbook unit would improve. This paper describes the design, development, and initial beta testing of

Table 1. Views of Diagnostic Assessment by Assessment Specialists

Author(s)	Explanation/Illustration of Diagnostic Assessment	Explanation of Relations with Other Types of Assessment
Alderson & Huhta 2005	DIALANG... "oriented towards diagnosing language skills and providing feedback to learners rather than certifying their proficiency" (p. 301)... "It not only reports to the learners, at different levels of detail how they did on the test" (table of right and wrong answers grouped according to subskill to give the user an idea of their strongest and weakest subskills), "but also gives them advice and awareness-raising information that is aimed at setting in motion a process of further language learning" (p. 305)	
Bachman, & Palmer 1996	"Diagnosis involves identifying specific areas of strength or weakness in language ability... ...so as to assign students to specific courses or learning activities" (p. 98))	"Illustrated by a test whose purpose it is to make decisions about whether or not job applicants ...who have been admitted to the course have mastered specific course content...used to provide diagnostic feedback to those who have not mastered the content" (p. 291)
Chapelle & Douglas 2006	"..teachers have access to computer-assisted language tests that are included as part of online language courses...early example was the French curriculum on PLATO...which kept records on learners' performance during each session of their work over the course of a semester... (Marty, 1981)" (p. 4)	"As Clark (1989) pointed out, diagnostic tests are developed according to different specifications from those used to construct a proficiency test"...achievements designed to match courses...developed through ...application of criterion-referenced testing (p. 5)
Davies, Brown, Elder, Hill, Lumley, & McNamara 1999	"used to identify test takers strengths and weaknesses, by testing what they know or do not know in a language, or what skills they have or do not have. Information obtained from such tests is useful for ...identifying areas where remedial instruction is necessary" (p. 43)	"Relatively few tests are designed specifically for diagnostic purposes. A frequent alternative is to use achievement or proficiency tests (which typically provide only very general information)." (p. 43)
Linn & Miller, 2005	"Persistent learning difficulties may require the use of diagnostic tests. For this type of testing, a number of test items are needed in each specific area, with some slight variation from item to item." (p. 136) "It is concerned with the persistent or recurring learning difficulties that are left unresolved by the standard corrective prescriptions of formative assessment" (p. 36). "Diagnostic testing is a highly specialized area that has been somewhat neglected in education ...Teachers have to depend more heavily on the diagnostic features of formative tests..." (p. 136)	Formative assessment... "typically cover some predefined segment of instruction (e.g., a chapter or particular set of skills) and thus encompass a rather limited set of learning outcomes...when a small number of students perform in ways that show a lack of understanding of critical concepts, alternative methods of study may be prescribed. Tests and assessments may be given at the beginning of an instructional segment to determine whether students have the prerequisite skills needed for the instruction (to determine readiness) ...Readiness pretest are typically limited in scope...tend to have a relatively low level of difficulty...and serve as a basis for remediation" (pp. 135-136).
Richards & Schmidt 2002	"a test that is designed to provide information about L2 learners' strengths and weaknesses. For example, a diagnostic pronunciation test may be used to measure the L2 learners' pronunciation of English sounds. It would show which sounds L2 learners are and are not able to pronounce or whether pronunciation is intelligible or not" (p. 155).	"Diagnostic tests may be used to find out how much L2 learners know before beginning a language course to better provide an efficient and effective course of instruction" (p. 155).

a prototype Readiness Check and Achievement Test for MEL.

Development of the Assessments

The *NorthStar* Reading and Writing book, Intermediate, Unit 6 was selected as the target unit for prototyping this project. This unit focused on the theme, Ecotourism.

***Readiness Check*.** The purpose of the Readiness Check was to determine whether the students were ready to start the unit and if not, to help them by diagnosing their weaknesses and providing materials that addressed those weaknesses. Specifically, the Readiness Check was designed to assess students' knowledge of vocabulary and grammar, to prepare students for in-class work, and to offer additional help to students who performed below a set criterion level. The Readiness Check and its associated remediation materials were created in five steps: (a) trialing Unit 6 materials with students, (b) preparing a content analysis, (c) designing task descriptions, (d) developing the Readiness Check, and (e) designing and developing remediation materials.

In the trialing stage, we worked individually with six students at the Intensive English Orientation Program at Iowa State University on Unit 6 materials. Students were asked about Unit 6 vocabulary, the meaning and pronunciation of words, problems they encountered when reading and listening to the unit's texts, and their difficulties understanding the content. Then we produced a summary of students' performances focusing on problematic language and content areas. In the second stage, a content analysis of the unit was completed by listing grammar points and lexical items which were not explicitly taught in the unit but which students struggled with. Since students did not seem to have problems understanding the subject matter (i.e., the content), it was not included in the Readiness Check. Then, the task descriptions were written and the test was developed.

As shown in Table 2, the Readiness Check contained one reading text and four activities: Vocabulary 1 tested word meanings (10 items), Vocabulary 2 tested word collocations and use (3 items), Vocabulary and Grammar tested sentence meanings (4 items), and Grammar tested the relationship between words and ideas (4 items). All of the exercises were variations on the alternative response or multiple choice item types with three or four choices except for the Grammar activity in which students chose the answer by clicking on a word/phrase in the sentence. Once students submitted their answers for each section, they received a score along with immediate feedback for incorrect answers consisting of an explanation of each correct answer. Figure 1 shows a Vocabulary 1 item with immediate feedback and score.

In the final stage, we designed and developed remediation materials that allowed for an adaptive instructional path to be taken by students who scored less than 80% on any of the four activities. These materials gave short explanations of vocabulary and grammar points and provided additional practice. Figures 2 and 3 illustrate the remediation materials for Vocabulary 1, which were automatically assigned.

Table 2. Task Descriptions for the Readiness Check

Text	• 1 text with similar topics and of similar difficulty as the unit text • Should include vocabulary and grammar points from the unit but not those explicitly taught (as identified in the content analysis). For example, words used to introduce new vocabulary, grammar constructions and sentence structure in unit texts.
Activity A Vocabulary 1	***Focus on word meaning*** • 10 words from text • Students check words they know (radio buttons: Yes- I know the word; No-I don't; I'm not sure) • For words checked they choose the correct meaning in the second part of the exercise (multiple choice with three answers)
Activity B Vocabulary 2	***Focus on word collocations and their use*** • 3 sentences • Sentences contain word collocations which are underlined. For example, an adjective from the text and a noun. There may be more than one collocation per sentence. • There are two or more collocations that can be used to replace the underlined one. Students click on correct collocations (multiple choice).
Activity C Vocabulary & Grammar	***Focus on sentence meaning*** • 4 sentences • Sentences contain difficult expressions, vocabulary, grammar constructions and their combination. For example, a tentative verb with a relative clause or a modal with a metaphor. • Students choose two sentences with the same meaning (multiple choice with four answers or more).
Activity D Grammar	***Focus on relationships between words or ideas*** • 4 sentences • Sentences contain underlined grammar constructions or word combinations. For example, relative clauses and verb-adverb combinations. • Students click to highlight the word(s) that refer to the underlined words.

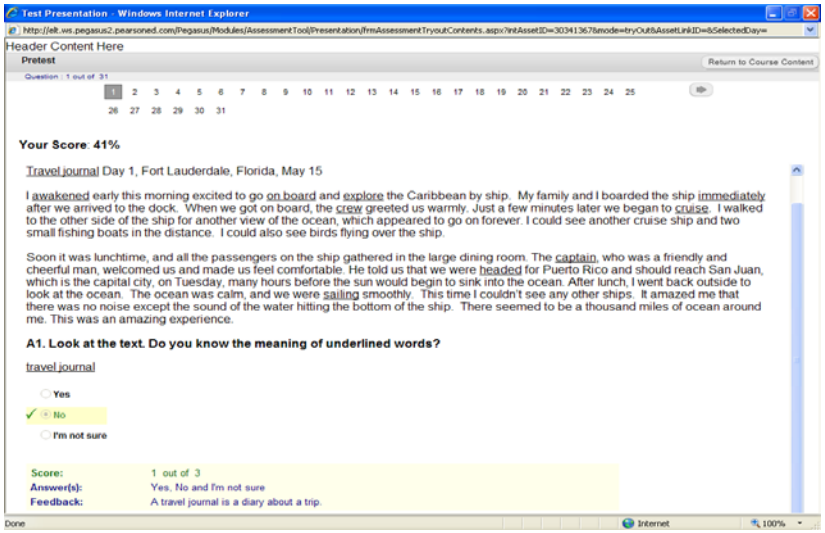


Figure 1. Vocabulary 1 activity in the Readiness Check with immediate feedback and score

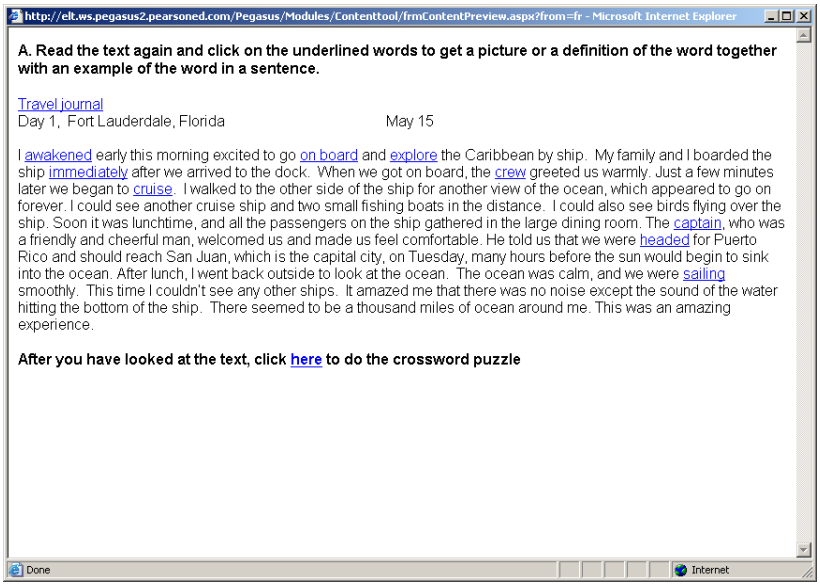


Figure 2. Vocabulary 1 remediation materials: Text with target lexical items underlined

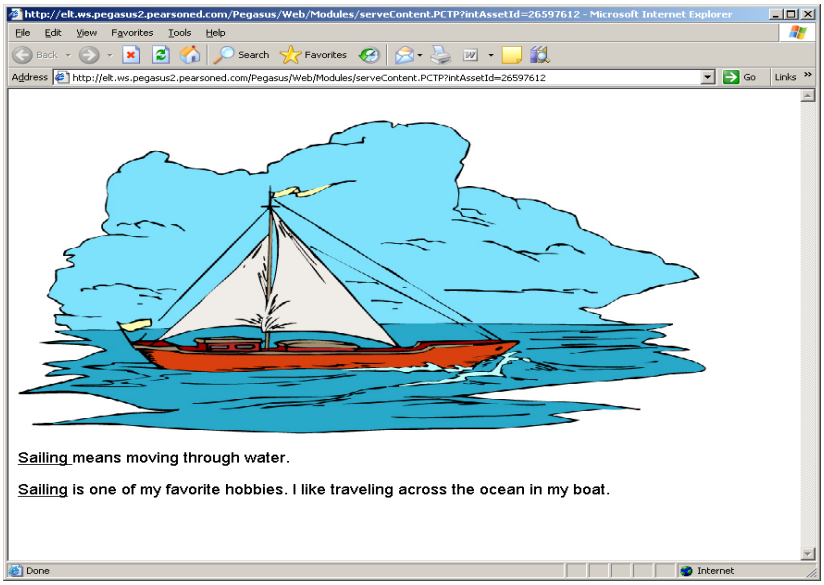


Figure 3. Vocabulary 1 remediation materials: Visual and textual annotation of a target word

In Figure 2, the key words are highlighted, inviting students to click on them. (Note also that a crossword puzzle was available for additional practice.) In Figure 3, we can see the result of having clicked on *sailing*—along with a picture, a definition is provided, and the keyword is used in a sentence.

Achievement Test. The development of the Achievement Test and its remediation materials was conducted in five stages: (a) scanning the scope and sequence document in the NS makes, (b) conducting the content analysis, (c) designing the table of specifications, (d) developing the test, and (e) designing and developing the on-line remediation materials.

First, the types of tasks and skills utilized in *NorthStar* were identified in the scope and sequence sections of the five levels of the Reading and Writing books. Within each Reading and Writing book, the tasks and skills were categorized into five sections: (a) critical thinking skills, (b) reading tasks, (c) vocabulary, (d) writing tasks, and (e) grammar. These tasks and skills provided the overall framework for the achievement test and the remediation materials. Following the scope and sequence review, an analysis of the unit content was then conducted.

The content analysis required that we identify the specific types of tasks and skills covered in the *NorthStar* Reading and Writing, Intermediate, Unit 6 book. Table 3 provides a summary of the content analysis for this particular unit. In doing the content analysis, we identified the tasks and skills, as well as the approximate number of tasks associated with each skill throughout the entire unit. Furthermore, the cumulative percentages of tasks and skills were calculated in order to identify the distribution of items found throughout the unit.

Table 3. Summary of Content Analysis

	Critical Thinking	Reading Skills	Writing Skills	Vocabulary	Grammar	# of Tasks	% of Tasks
Part 1. Vocabulary							
Vocab				10 definition; 19 analogy; 8 word association		37	34%
Part 2. Reading							
Reading #1 Reading #2	5 inference	4 prediction; 3 id chronol. order; 12 details				24	22%
Integrate readings	3 compari-son	5 write dialog paraphrasing opinions from 2 readings				8	7%
Part 3. Writing							
Writing: Edit, copy, sentence, outline			3 edit; 6 copy; 16 outline	5 write sentences	11 sentences with grammar: because/even though	41	37%
Writing: Essay	point of view		opinion			1	N/A
# of Tasks - Essay	8	19+5	25	42	11	110	
% of Tasks	7%	22%	23%	38%	10%		100
# of Writing Tasks			1			1	
% of Writing Tasks			100				100

Note. Other integrated items, speaking = 13; experiential = 4; opinions = 3; main idea and details = 4; discussion of editing =3.

For example, there were a total of 42 vocabulary items in the unit, which comprised approximately 38 percent of all items found in the unit. A list of integrated items was also noted, but these items were not included in the actual content analysis summary. Once the content analysis was completed, the Table of Specifications for the Achievement Test was designed.

Table 4 provides an overview of the Table of Specifications (TOS) used for the Achievement Test. A TOS is important because it specifies what will be included in a complete test. It is a two-way chart which outlines course content and objectives along x and y axes, assigning weights that reflect their relative importance in the syllabus. In this way, it provides a framework for measuring what was taught (Linn & Miller, 2005; Stoynoff & Chapelle, 2005).

As indicated in the TOS, the skills and abilities identified in the scope and sequence of the textbook are listed in the top row of the chart (the x axis). Looking down the first column on the left (the y axis), one can see that the achievement test was divided into two parts, Reading and Writing, each with two subsections included within them (i.e., reading & integrate readings, and editing/revising writing & writing). Within the cells are the subskills (e.g., inference, predict), test sections (e.g., 1.1, 2.2) and the number of test items included in the Achievement Test. The final pilot version of the Achievement Test consisted of 43 items worth a total of 52 points. The item types used in the test included

Table 4. Achievement Test's Table of Specifications

	Critical Thinking Skills	Reading Skills	Vocabulary	Writing Skills	Grammar	# of Tasks/ Points	% of total points on test
Part 1. Reading							
Reading (15 min)	2 1.4(2) Inference	9 1.1(1) Predict 1.2(4) Details 1.3(4) Details	12 1.6(6) Definitions 1.7(6) Analogies			23	44%
Integrate readings (5 min)	8 1.5(8) Comp/Contrast					8	16%
Part 2. Writing							
Editing/Revising writing (10 min)				7 2.1(1) Identify 2.2(3) Organize 2.3(3) Edit	4 2.4(4) Cause/Effect	11	21%
Writing (20 min)	(see opinion essay)			1 2.5 (1) opinion essay (10 points)		1 (10 points)	19%
Total points on test	10	9	12	17	4	52	
% of total points on test	19%	17%	23%	33%	8%		100

multiple choice, matching, fill-in-the-blank (42 items scored dichotomously, worth 42 points), and essay writing (one essay worth 10 points). Note that the percentage distributions of the skills/abilities on the test are comparable to those in the textbook.

For the Achievement Test, the passing scores were set at 75% for each of the five skill/ability sections. If students obtained scores below this, they were asked to complete the on-line remediation materials that supplemented each of the five sections. The remediation materials included the following components: (a) making inferences from the reading, (b) organizing the opinion essay, (c) writing better opinion statements, (d) vocabulary in the unit, and (e) grammar in the unit. Each component consisted of tutorial-style exercises that aimed at guiding students through potential problem areas encountered throughout the unit and during the Achievement Test. The tutorials typically began with explanations and guided practice exercises, and were followed by exercises in which students provided answers. Correct and incorrect answers received item-level feedback, and students were encouraged to return to items they answered incorrectly.

NorthStar Pilot Study

Once all of the prototype materials were published in MEL, the beta test, described here as the "pilot study" started in July, 2007.

Participants. An intact class from an intensive English program at a public university in the US participated in the pilot. This program was selected because it used *NorthStar* textbooks and because Pearson Longman, the publisher of the series, had successfully collaborated with the program director and teachers in the past. A teacher who taught using the *NorthStar* textbook for Intermediate level was contacted and agreed to participate in testing new Unit 6 paper-based materials and online MEL materials. The teacher had 20 years of experience teaching ESL and had worked for the ALP program for 9 years.

There were 13 students in the group who signed the consent form. The great majority (10 students) of students were female. Students came from a variety of countries, as shown in Table 5, though the majority were from Korea.

Textbook materials. The *NorthStar* textbook materials covered in each class and assigned for homework are presented in Table 6. The class spent 3 class periods on Unit 6. The class did not cover predicting, editing, or topic expansion, but did cover most of the Unit 6 materials in class or for homework. The class spent class time on sharing information, analyzing the structure of the essay, and outlining an essay (though students did not write an essay).

In addition to Unit 6 textbook materials, students were assigned MEL materials for homework. The class was assigned vocabulary, reading, grammar and writing activities in MEL in addition to the Readiness Check and the Achievement Test. The students were assigned MEL homework after each class period in addition to the regular homework. The teacher assigned all available activities for the four language areas—vocabulary,

Table 5. Students in the Pilot Study

ID	Gender	Country	ALP placement test	Data taken	Placement
21	M	UAE	38	Summer A 07	3a
22	F	Korea	no score reported		3b
23	F	Korea	no score reported		3b
25	F	Italy	19	Summer A 07	1
26	M	Turkey	27	Summer A 07	2a
27	M	Azerbajan	did not test		
29	F	Korea	39	Summer A 07	2a
30	F	Korea	42	Summer A 07	3a
31	F	Korea	46		3b
33	F	Korea	27	Summer A 07	2a
34	F	Korea	33	Spring 07	2b
35	F	Korea	42	Summer A 07	3a
36	F	Turkey	28	Spring 07	2a

reading, grammar, and writing—but did not assign the background map activity, or the internet, video, and integrated task activities.

Procedure. The teacher was asked to plan about 4 class periods (1hour 50 minutes each) for the unit together with the Achievement Test. She agreed to be observed teaching during one of her classes, to be interviewed twice out of class, and to fill-in a teacher questionnaire. In addition, the teacher agreed to have students do two out-of-class group interviews and fill in a student questionnaire. She was given 30 minutes of hands-on MEL training, but did not participate in the one-hour on-line Webex presentation of MEL before she started the unit. One of the authors visited her class and demonstrated MEL to students in 20 minutes. Students were assigned the Readiness Check for homework after the 1st class period. The second class period was observed and students were interviewed after class. Before the 3rd class period, the teacher was interviewed and the students were interviewed after that class. The teacher also completed the teacher questionnaire. Two of the 13 students had data saved from the Readiness Check. Nine of the 13 students took parts of the Achievement Test. Seven students completed parts of the Achievement Test and student questionnaires in MEL. Two students who did not do the Achievement Test in MEL were administered the paper-based version of it (three other students took parts of the Achievement Test on paper).

Results. At the end of the pilot study, student participants were given questionnaires and were interviewed in focus groups. Two of the thirteen students apparently took the Readiness Check. Based on the set passing score of over 80%, both students should have

Table 6. Unit 6 Materials and Class Coverage

Textbook Content of NS RW Ch. 6	Class period	Minutes
1 Focus on the Topic		
A Predicting		
B Sharing Information	1	35
C Background and Vocabulary	1, 2	5 (no vocab); homework; 5
Antarctica Quiz		
2 Focus on Reading		
A Reading One: Tourists in a Fragile Land	1	homework
Reading for main ideas	1, 2	homework;6
Reading for details	1, 2	homework;6
Making inferences	1, 2	homework;18
Expressing opinions	2	6
B Reading Two: A Travel Journal	2	17
Discussion questions	2	18
C Integrating Readings One & Two	2	
Organizing	2	7
Synthesizing	2	13
3 Focus on Writing		
A. Vocabulary		
Review		5
Expand	2,3	homework, 10
Create		optional homework
B Grammar for Writing	2,3	homework, 10
C Focused Writing Task		
Preparing to write	3	
Organizing		100
Revising	3	20
Editing		
D Topic Expansion		
Writing		
Research		

Table 7. Student Responses - Readiness Check (N=6)

Question	Strongly agree	Agree	Disagree	Strongly disagree	No opinion
The Readiness Check was helpful.		4	1		1
The Readiness Check was easy.		4	1		1
The recommended study materials in the Readiness Check helped me to understand the vocabulary and grammar in Unit 6.		5	1		
I liked the feedback for wrong answers in the Readiness Check.		4	2		

been assigned and should have completed three remediation activities. However, only one student did one activity. Two possible explanations why there were only two student records are technical in nature. First, the Readiness Check could be taken only once and the students who opened it to see what it was like could not get back to it. Second, MEL did not keep records of students who did the Readiness Check after the due date. The Readiness Check was due on the day when the class did not meet but some students took it on a later day. In this case, the students could take the test and were assigned remediation, but their score did not appear in the grade book. As a consequence, MEL did not save their submissions and it appeared that the students did not take the Readiness Check at all. Based on the questionnaires and interviews, it seems that perhaps six students did take the Readiness Check.

According to questionnaire responses, most of the students had favorable comments about the Readiness Check. Four out of six students agreed that the Readiness Check was helpful (see Table 7). Four students found the Readiness Check easy, which is the response we anticipated because we wanted most of the students to do well on the Readiness check. Also, five students agreed that the Readiness Check and study materials helped them understand the vocabulary and grammar in Unit 6. Finally, four students liked the feedback they received for wrong answers.

In the interview, one student commented that she knew all vocabulary she was tested on; nevertheless, this student found the Readiness Check helpful because the style of presentation was different from the classroom. One less positive comment concerned Activity C. Students mentioned that they found Activity C difficult and confusing because they were asked to select two correct answers. This type of response was different from Activity A where there was only one correct answer in multiple choice questions. This apparent confusion could be addressed through different direction lines, or by changing the number of correct responses.

The teacher was also interviewed and responded to a questionnaire. The teacher liked the Readiness Check, found it useful, and commented that “it sensitizes students to the material before they actually get to them.”

Seven of the thirteen students took parts of the Achievement Test online; two others took it on paper. Seven of the nine students who took the test did not meet the 75 percent criteria for passing in at least one section. Based on the passing score of 75% or higher, students seem to have been assigned and did work on the appropriate recommended study materials. Table 8 indicates these students' performances for the different skills sections in the on-line test administration.

Scores and remediation results could not be reported for students' essays since this test section was not assigned by the teacher in the pilot study. Meanwhile, Table 9 indicates that six of the seven students worked on the suggested on-line remediation materials. This is an encouraging finding because students followed the adaptive path recommended in MEL without the teacher having to assign extra work.

Table 8. Percentage Scores on On-line Achievement Test's Sections

Student	Reading & Critical Thinking	Vocabulary	Editing	Grammar	Essay
Student 22		50	75	100	
Student 23			38	100	
Student 25	58	75	88	100	
Student 29		92	13	100	
Student 30		83	63	100	
Student 33	37	92	38	100	
Student 34	53	83	25	50	

Table 9. Remediation Activities based on Achievement Test Performance

Student	Reading & Critical Thinking	Vocabulary	Editing	Grammar	Essay
Student 22		X			
Student 23					
Student 25	X				
Student 29			X		
Student 30			X		
Student 33	X				
Student 34	X		X	X	

Note: X = recommended remediation material begun

Table 10. Student Responses – Achievement Test (N=6)

Question	Strongly agree	Agree	Disagree	Strongly disagree	No opinion
The questions on the Achievement Test were like the ones in the unit.		4	1		1
My Achievement Test score was fair.		3	3		
The recommended study Material for the Achievement Test provided good extra practice.		5	1		

As Table 10 shows, the majority of students agreed that the Achievement Test questions resembled those in the unit. Five students believed that the recommended study materials provided good extra practice. Only half of the students agreed that their score was fair.

This could be a cause for concern. Students 23, 33, and 34 did not think that their test score was fair. Further inspection revealed that these students received failing scores on half of the sections they took, but the other four students who took the test online also failed some sections. The three students who had negative impressions had not begun any of the other MEL activities. It is unfortunate that we could not probe more deeply into this issue. At the time the interview was conducted, only two students had done the Achievement Test and they could not give any feedback. We should continue to ask this question, observe results, and address any deficiencies in the test format or delivery.

Finally, on the questionnaire, the teacher had "no opinion" about how the students did on the assessments, or whether they did the remediation activities. This is unfortunate as the prototype materials are intended to help the students and the teacher. However, it is also understandable, as the time for the pilot was very brief, included only one unit, and the teacher had no prior experience with the MEL platform. The teacher did strongly agree that the questions on the Achievement Test were like those found in the unit. She also strongly agreed that automatic grading of the MEL activities and tests saved her significant time.

CONCLUSION

The assessments with their recommended study materials reflect the content that is in the textbook, providing more feedback and practice that is perceived as helpful. It appears that the students were generally excited about the materials and indicated that they helped them learn English. This opinion was shared by their teacher. At this time, the prototype Readiness Check and the Achievement Test promise a number of apparent advantages both for students and teachers. First, students' weak performances can be identified and individuals can be offered immediate scores and feedback. Second, learners can be given extra explanations and practice as a means of remediation. Students liked the immediate score reporting for both tests and believed that Achievement Test study materials provided good extra practice. Also, teachers benefit from automated grading of answers which saves time while automated record keeping allows easy monitoring of students' progress.

Whether these promises and others will be kept remains to be seen. We believe that immediate feedback and remediation is beneficial, but at the risk of stating the obvious, it is important that students actually take the assessments in order to receive the feedback and remediation. The data collected in the pilot were not sufficient to support benefits of the diagnostic use of these tests for *NorthStar*. Still, we consider this project to be a small step forward in implementing on-line diagnostic English language assessment, as an accompaniment to textbook-based instruction. In the future, it will be important to note whether students who do not perform well actually complete the remediation materials

and think that they are helpful for them. It will also be important to monitor teachers' involvement with the automated on-line materials, both in terms of assigning materials and monitoring students' progress. If this project is to succeed, the on-line materials must complement traditional textbook-based classroom instruction; they must not be an isolated add-on to it.

We hope to examine more data from student performance and questionnaire responses as they become available. This will allow us to probe more deeply into the effectiveness of using the results from readiness pretests and formative achievement tests as a basis for adaptive diagnostic assessment.

REFERENCES

Alderson, J. C., & Hakuta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22,* 301-320.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. New York: Oxford University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. New York: Cambridge University Press.

Clapham, C., & Corson, D. (Eds.) (1997). Language testing and assessment, volume 7. *Encyclopedia of Language and Education.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of language testing. *Studies in Language Testing 7,* Cambridge, UK: Cambridge University Press.

Linn, R. & Miller, M.D. (2005). *Measurement and evaluation in teaching*. (9th Ed.). Upper Saddle River, NJ: Merrill, Prentice Hall.

Richards, J., & Schmidt, R. (2002). *Longman dictionary of language teaching & applied linguistics*. New York: Pearson Longman.

Stoynoff, S., & Chapelle, C. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Zhang, Q. (2008). Towards cognitive response theory in diagnostic language assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 40-61). Ames, IA: Iowa State University.

Towards Cognitive Response Theory in Diagnostic Language Assessment

Quan Zhang
College of Foreign Studies,
Southern Medical University, P. R. China

This paper proposes that Cognitive Response Theory (CRT) be implemented in the form of computerized cognitive testing (CCT). It begins by contrasting key characteristics between traditional computer adaptive testing (CAT) and CCT. CCT is operationalized through a jumbled word (JW) test item, yielding two cognitive variables—response type and response time—to estimate ability level. The latent variables hypothesized to underlie test performance were tested against the data through the use of structural equation modeling (SEM). Results show promise for applying CRT to tests of English as a foreign language.

INTRODUCTION

The growing reliance on tests for making high stakes decisions and for improving educational outcomes has called attention to some serious limitations germane to theories guiding language testing practice.[i] In the author's view, scholars and experts of language testing should actively address such problems by refashioning assessments to meet current and future needs for quality information with the help of cognitive science as well as computer and multimedia technology. Take the multiple choice (MC) question format as an example. For many years, MC has been the dominant and indispensable test format in assessments ranging from teachers' informal quizzes to large-scale tests administered worldwide. Traditional computer-adaptive tests also use the MC question format to explore diagnostic assessment. Though the theoretical basis for the MC question format can be found, relevant criticisms have also been voiced.

This paper argues that with the advanced technology of computer programming and multimedia, a jumbled word (JW) item format is a promising alternative to the MC item format for language assessment. I begin by explaining some of the limiting characteristics of the MC item formats in contrast to those of the JW format. I then discuss some aspects of a pilot study that was conducted to evaluate an English grammar test in the JW format. The basis of the JW item test was based on Cognitive Response Theory (CRT) realized in the form of computerized cognitive testing (CCT). A CCT model is believed to better

The author would like to acknowledge with thanks Prof. Carol A. Chapelle and her organizing committee of the 5th annual TSLL conference, whose invaluable comments and suggestions for improvement helped refine the paper. The author's thanks also go to Eric Wu of Department of Psychology, UCLA, under whose guidance both CCT modeling and EQS running went well. The opinions expressed in this article are those of the author who remains solely responsible for any possible errors in the article. This research is supported by funding of Guangdong Provincial Project for New Century, PRC (2006-2008).

demonstrate the "Assessment Triangle"[ii] by explicitly connecting cognition, observation and interpretation. The research presented here, though preliminary, calls for feedback from the larger community of language testing. It is intended to demonstrate a tangible basis for further research towards diagnostic assessment via CCT.

THE TRADITIONAL COMPUTER ADAPTIVE TEST

Although the original idea of adaptivity can be traced back to the work of Binet (1909), Lord (1970), Birnbaum (1968), and others, only with the advent as well as availability of computers nowadays could the traditional computer adaptive testing (CAT) method become feasible for widespread operational research and implementation. "Adaptive" in this sense refers to a testing procedure that selects the next item to be presented to an examinee based on a test taker's performance on the previous one (Bunderson et al, 1989). Such a procedure is based on the idea that more information about a test taker's trait can be obtained from an item with a difficulty level fitting the test taker's ability. Therefore, an adaptive test requires a set of test items at various difficulty levels, nowadays mostly still in MC format, from an item bank. These test items are calibrated in advance so as to yield parameters that can be used by the selection procedure during test taking. CAT is widely acknowledged to have advantages over conventional paper-and-pencil tests in estimating test taker's ability; however, some characteristics of CAT limit the possibilities of measurement for language assessment. In light of cognitive theory, the following limitations are evident in current CAT practices: CAT is difficulty-bound, dichotomous-valued, MC-limited, time-neglected, and product-oriented (Zhang, 1993, 2002a).

Difficulty-bound refers to the fact that CAT relies on only one aspect of cognition, i.e., the adaptivity procedure is controlled by the item difficulty and a test taker's ability. In other words, ability is bound to item difficulty alone, each being interpreted only in the specific context of the other. The unidimensional ability-difficulty link is overly simplistic in view of the many cognitive variables that influence test performance to some extent. As a consequence, CAT cannot explicitly include substantially meaningful interpretation of what test performances should actually be inferred to mean.

CAT is dichotomous-valued in that it treats the adaptivity between item difficulty and a test taker's ability as binary-valued, i.e., yes-or-no type. Test takers who present wrong answers are all labeled as lacking certain knowledge in the tested domain. In this sense, CAT fails to distinguish test takers who have partial knowledge from those who don't have any in the process of problem-solving. This concerns very much test validity. In the view of cognitive science, human's cognitive ability is by no means dichotomous-valued. Instead, it is of more-or-less type. Over the past 40 years many researchers (Rasch,1960; Bock, 1981; Hambleton, 1989; Hambleton & Swaminathan, 1985; Smith, 1987; Mislevy & Bock, 1984; Mislevy & Verhelst, 1987) have examined the hypothesis that dichotomous scoring does not capture the full information available in the responses concerning a person's cognitive ability. Most of these scholars have found that the degree of incorrectness of an answer can be quantified and used as an additional source of

information about the test taker's ability. To overcome this perceived deficiency of dichotomous scoring, a variety of techniques have been developed, such as response weighting, answer until correct, degree of confidence weighting, elimination scoring, and so forth (Smith, 1987).

The current CAT practices are mostly limited to a MC question format despite the fact that at least four problems with this item format have been found. First, good MC items are difficult to develop. The common practice is that each question stem and distractor, prior to its use, undergoes the process of item writers' moderation and pretesting. Second, tackling a MC question is a selective process rather than a precise one. It involves partial use of available minimal language cues selected from perceptual input on the basis of the test takers' expectation. As this partial information is processed, tentative decisions are made to be confirmed, rejected, or re-confirmed as coping with each item progresses (Snow & Lohman, 1989). Third, the attempts at the distractors made by test takers may reveal their cognitive level. In other words, the possible guessing behavior demonstrated in selecting the distractors may be taken as the indicator of test takers' cognitive ability. Ideally, such data should be utilized by test developers in post-test item analysis to gain insight into the test takers' cognitive ability. Finally, with further understanding of cognition, test users have also come to realize the importance of the guessing factors. Overall, it seems evident that MC question format should by no means be the only test form for measuring cognitive abilities, and language ability in particular.

CAT is also time-neglected as it does not record test takers' reaction time during test taking. The failure to collect these data prevents CAT from distinguishing among the abilities of test takers who obtain the same scores. This again concerns test validity because a single score on a test can be obtained in different ways by different examinees. In the view of cognitive science, solution time is an important cognitive variable particularly in measuring the procedural knowledge at command. It distinguishes experts' from novices' performances. Hence, without tracking test takers' solution time, CAT practice might in some way weaken the test validity.

When it is said that CAT is product-oriented, it means that CAT does not assess test takers' problem-solving strategies or skills. What a score from CAT reflects is just the terminal answers, which are either right or wrong, and which offer no chance for test users to observe examinees' problem-solving procedures. In the view of cognitive science, the significant interpretation of test takers' real potentiality pertaining to problem-solving is evident from a display of their cognitive process rather than the terminal product. Therefore, the meaningfulness of inferences drawn from CAT assessment using MC questions may be compromised despite the fact that today's multimedia and web technologies make such observations feasible.

COMPUTERIZED COGNITIVE TESTING

Computerized Cognitive Testing (CCT) is a theoretical approach that is intended to provide an alternative to traditional CAT. It is therefore useful to examine its potential to

be applied in language assessment (Zhang 1993, 2007). Compared with CAT, CCT is unique in the following six aspects: CCT is cue-provided, polychotomous-valued, JW-adopted, time-recorded, process-oriented, and procedural-knowledge-based. Such features of CCT will be explained in detail in this section to show its advantages over traditional CAT.

CCT is cue-provided as it is capable of giving hints relevant to the solution to a problem in case that test takers fail to provide a correct answer at the first attempt. Evidence for the importance of providing such cues or hints comes from experiments showing that most students who failed to provide a correct answer at the first trial were not ignorant of or lacking in the relevant knowledge. Given a little help, they could quickly solve the problem. Then, why is the provision of hints better than difficulty-based adaptivity? In traditional CAT, an examinee will be given a different test item at a lower difficulty level if he or she previously fails in a relatively more difficult item. However, two items at different difficulty levels are usually of different content and may thus test different aspects of language ability. So the construct in such a test is inconsistent across items or defined in a way that encompasses the content of all the items. In contrast, CCT provides hints to lower the difficulty level of a test item but still keeps its content unchanged. In other words, both the difficult and the easier items assess the same language aspect, reinforcing the diagnostic assessment.

How to provide students with immediate and direct feedback on their test performance is one of the problems in education. In China, whether the test is a placement test, an achievement test, or a proficiency test, students usually do not receive feedback about their test performance until a long time after they have taken the test. Furthermore, what they receive is usually general feedback, such as information about what is the best choice for a multiple choice question. This is not only because we lack certain research methods for monitoring students' performance during a language test but also because it is not feasible to implement such a task in any traditional paper-and-pencil tests.

According to Anderson (1974, 1976, 1983, 1985), the retrieval process in information processing requires that certain cues be provided, either by the external stimulus or by the learner. Accordingly, developers of diagnostic assessment must consider providing such cues. This is based on a cognitive hypothesis that human knowledge is stored by means of propositional networks in the brain. The retrieval of the knowledge from the brain is achieved through the spread of activation. Thus, it appears that a well-organized knowledge structure in the brain is activated more quickly than otherwise. In some cases, the retrieval process may be baffled due to the lack of certain knowledge. Therefore, according to CCT a test taker failing to provide a correct answer for the first time may have the relevant knowledge but be unable to activate it due to the inefficient organization of the knowledge in the brain. In this case, a certain external stimulus is required to trigger the retrieval of that knowledge. In this sense, the significance of providing cues is twofold: A language test with cues is capable of (1) distinguishing test takers who have the relevant knowledge stored in the brain from those who are totally lacking in such knowledge and (2) further discriminating test takers who answer the same

number of questions correctly by revealing how they get the answers right. Hence, the provision of cues is potentially an important approach to diagnose a test taker's real language ability.

"Polychotomous-valued" means that CCT treats the 'cognition' between item difficulty and an examinee's ability as a continuum in terms of the degree of achievement. CCT presumes that test takers who fail on a test item for the first time may possess partial knowledge in the tested domain. In contrast to CAT, it offers hints to the answer so as to give the test taker more chances to succeed. If the test taker gets the right answer with the help of the hints, CCT also gives a partial credit. In this way, CCT can further distinguish test takers who have given the same number of correct answers based on the extent to which they seek help. The range of ability levels thus interpreted by CCT goes from the highest to the lowest with many intermediate scores in between. Such a detailed ability continuum, which maps observed responses to the strength of knowledge, best reflects the different levels of human cognitive ability, and thus proves to be another way to explore the real language ability of test takers.

The JW task form used by CCT has three advantages. First, the JW form allows for the assessment of integrative skills concerning both vocabulary and grammar knowledge. Second, the JW form demands dynamic performance. The third advantage of the JW form is that it prevents test takers from making a blind guess about the correct answer because the available language cues for guessing are minimized. In addition, the JW form is intended to assess test takers' procedural knowledge (knowing how) as well as declarative knowledge (knowing what) in cognition. The cognitive basis of the JW task design can be verified in the following three aspects:

- The JW form focuses more on the use of language than on the knowledge of language per se. In other words, it focuses more on procedural knowledge than declarative knowledge. In the view of cognition, procedural knowledge entails declarative knowledge. Thus, the integrative skills concerning both vocabulary and grammar knowledge assessed via JW test form is in fact test takers' procedural knowledge. In this sense, a test taker's quick, correct response to a JW item without the assistance of cues reflects his/her solid possession of both declarative and procedural knowledge, while a correct answer based on cues reveals that the test taker has the knowledge but that it has not been proceduralized.
- The JW test item demonstrates that the whole is more than the sum of its parts, an important Gestalt claim. In this sense, one's language ability is by no means merely the sum of one's vocabulary and grammar knowledge put together. This can be justified by the experiment and post-experiment interviews (see discussion of the experiment in the following section) conducted by the researcher. The interviews reveal that that some subjects who knew the meanings of individual jumbled words were still unable to put them into a logical sequence. Besides, we found that not all the subjects who knew the concept of tenses or

attributive clauses could fulfill the tasks well. In other cases, test takers could quickly get the correct answer when they used specific hints given by the test. Cognitively, this is interpreted as such that the test taker's declarative knowledge not being very well proceduralized.

- The JW test item, in contrast to the MC question form, allows for no random guessing. In cognitive science, the extent of guessing indicates a person's level of cognition. In other words, the guessing behavior demonstrated in coping with JW test items is considered as the indicator of test takers' language ability. Thus, a correct answer obtained through guessing indicates the test taker's knowledge concerning the subject-matter learning is incomplete. Similarly, failure to obtain a correct answer after consulting hints is interpreted as total lack of the relevant declarative knowledge being tested.

CCT is time-recorded. Another method of evaluating cognitive processing is to measure the amount of time test takers spend in problem solving (Klahr & Robinson, 1981; Anderson & Gluck, 2001)[iii]. Data collected in this way can be highly informative. Here, time-recording refers to the reaction time or retrieval time spent by test takers in coping with each set of jumbled words during test taking. According to the speed at which a problem is solved or, in other words, certain knowledge is activated, CCT could produce a time parameter indicating test takers' levels of proficiency in six categories: Native User, Near Native User, Good User, Modest User, Average User and Poor User,. As noted previously, procedural knowledge is executed rapidly and with minimal demands on attentional resources; therefore, assessment must take into consideration the solution time, which is one useful index of automaticity for many problem-solving tasks. This has been proven to be another "window on the mind" (Dillon, 1985; Just & Carpenter, 1992)[iv] to observe the strategies test takers use in coping with JW test items.

When we say that CCT is process-oriented, we should first spotlight "process" in the sense of testing. For instance, in any tests of mathematics or geometry, test takers are usually required to write down the process to obtain a correct answer because each step in problem-solving indicates his/her relevant knowledge pertaining to the subject-matter learning. However, such an important cognitive assumption has not been paid much attention to in many language tests, particularly in the tests composed of multiple-choice questions. This is largely due to two reasons. On the one hand, test developers lack certain research methods in observing or measuring test takers' cognitive process during test taking. On the other hand, it is not feasible to do so in any traditional paper-and-pencil tests.

With the advances in computer technologies and the availability of computers, it is now feasible to make CCT process-oriented. Built on an information processing model, CCT is able to trace test takers' cognitive process of problem solving by recording their reaction time and remembering their use of the help options (i.e., cues). Thus, CCT is concerned more about how test takers get to the answer than whether the answer is right or wrong. In this sense, CCT can be considered an instrument to observe and trace what

is going on inside the test taker's brain while he or she is tackling each test item. This is what "process" means in the sense of cognitive testing, and the idea of assessing testing processes is in the spirit of "Assessment Triangle" as described in *Know What Students Know* (National Research Council, 2001).

Test takers' procedural knowledge can be best assessed by letting them arrange and re-arrange jumbled words into a logical sentence. This requires them to demonstrate how they sequence ideas in a second language. In this sense, the relevant procedural knowledge of sentence formation is tested. In view of cognition, good procedural knowledge entails good declarative knowledge. CCT implemented in this research

Table 1. Summary of Contrasts between CAT and CCT

Traditional Computer Adaptive Testing (CAT)	Computerized Cognitive Testing (CCT)
Difficulty-bound only Adaptivity is realized by adjusting item difficulty based on test takers' responses Difficulty is reduced by providing easier items, which, makes the language aspect being tested inconsistent.	**Cue provided** Two factors are considered for adaptivity: (1) Response type (2) Response time Difficulty is reduced by providing relevant hints and keeping the language aspect being tested unchanged.
Dichotomous-valued Uses binary logistic model, i.e. yes-or-no type rating. All the incorrect answers are treated as wrong. Ability estimation is bound to item difficulty alone.	**Polychotomous-valued** Uses partial credit model, i.e. more-or-less type rating All the correct answers are specified in different response types. Ability is estimated based on a continuum of degree of achievement.
MC-limited Requires separate skills only; Demands selective performance only; Offers chances for blind guessing.	**JW-adopted** Requires the integrative skills concerning both vocabulary and grammar knowledge. Demands dynamic performance. Offers no opportunity for blind guessing
Time-neglected Test takers' reaction/solution time is not taken into consideration for ability estimation.	**Time-recorded** Response time is taken as important cognitive variables for ability estimation
Product-oriented Test takers' strategies or skills demonstrated during problem-solving are not traced. Only terminal answers are scored.	**Process-oriented** Test takers' strategies or skills demonstrated during problem-solving are recorded for further diagnosis.
Declarative-Knowledge-based Reflects only declarative knowledge.	**Procedural-Knowledge-based** **Procedural knowledge is tested.** Best reflected in arranging and re-arranging a set of jumbled words into a logical sentence; Best demonstrates how integrated knowledge of language works independently; Good procedural knowledge entails good declarative knowledge.

generates a file for each test taker, recording the whole process of problem solving. This includes the test taker's response using or not using the hints, the item difficulty, the corresponding solution time, and the frequency of attempts. From such records, the test user is able to know exactly how the test taker solves each problem and how and why he or she fails. By tracking the test taker's behavior, CCT can find out his or her problems in information processing so as to provide individualized feedback for the follow-up instruction.

A Summary of the Differences between CAT and CCT

The six contrasting aspects between traditional CAT and CCT are summarized in Table 1. It highlights the important differences in measurement between these two testing approaches. Such differences shed light on the development of tests that can accurately measure examinees' knowledge and skills as well as on the development of tests designed for specific purposes such as diagnosis.

THE PILOT RESEARCH

The remaining part of the paper will report a pilot study which investigated the use of CCT for assessing examinees' English language ability. The study revealed the potential for CCT to be applied in language assessment. The methodology for the pilot study is built upon other research that the author has been conducting since 1993.

Partcipants

The original sample of participants consisted of approximately 200 vocational students in majors other than English at Southern Medical University, Guangzhou, China. The post-test data editing confirmed that 120 cases were valid data records.

Table 2. 15 Possible Response Types for CCT

Type	R/W	Hint-1	R/W	Hint-2	R/W	Hint-3	R/W
A	1						
B	0	N	1				
C	0	N	0	N	1		
D	0	N	0	N	0	N	1
E	0	Y	1				
F	0	N	0	Y	1		
G	0	N	0	N	0	Y	1
H	0	Y	0	N	1		
I	0	Y	0	Y	1		
J	0	Y	0	N	0	N	1
K	0	Y	0	Y	0	N	1
L	0	Y	0	Y	0	Y	1
M	0	Y	0	N	0	Y	1
N	0	N	0	Y	0	Y	1
O	0	N	0	Y	0	N	1
W							

Test Design

Ten JW test items (see Appendix A) were designed, each including 3 relevant hints. The average number of words used for each JW item was seven. Table 2 illustrates all the possible response types for the test. In the table, 1 and 0 indicate correct and incorrect answers while Y and N indicate whether or not a test taker used a specific hint. As each JW test item has three relevant hints, there are a total of 15 response types.

Data Analysis

The data of test takers' responses were analyzed using PARSCALE4.1. (See Appendix B for the PARSCALE command file.) Technically, PARSCALE only accepts ordinal data; therefore, the interval data of test takers' response time were coded into six ordinal categories: (1) Native User, (2) Near Native User, (3) Good User, (4) Modest User, (5) Average User and (6) Poor User. In addition, test takers' response times on each item were also recorded. The ability scores and the categorical response data with response time assume the partial credit model with the standard scoring function. In sum, the assessment of test takers' ability took two cognitive variables into consideration: response type and response time.

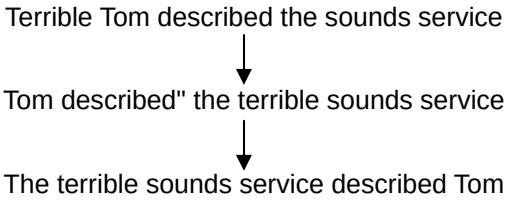
Two examples provided here are test takers' responses on two test items in CCT. It is worth noticing that the response types labeled 'E', 'F', and 'I' (See Table 2) are typically syntactic-knowledge based, or rather procedural-knowledge based. Accordingly, these examples best illustrate how participants approached the jumbled word items with the help of hints provided.

Example One

Subjects were presented a set of jumbled words as follows:

terrible, Tom, described, the, service, sounds

They were unable to identify the hidden syntactical structure of the target sentence at the first trial. So the first attempts they made were sentences such as:



Receiving such responses, a CAT system would presume the test takers are unable to cope with such an item and thus provide them with an easier one. As a result, the test item is changed and meanwhile, the language aspect being tested will probably be changed as well. In contrast, a CCT system would react differently in such a situation. It would provide a relevant hint germane to the item instead of providing a new test item. In this

example, the first hint the system provided is "*The sentence Begins with 'The service'*"; the second hint is "*This sentence contains an attribute clause*". These hints make the item easier for test takers to tackle. With the hints provided, the subjects appeared to become aware of the existence of an imbedded attribute clause and made a complex sentence as follows:

The service Tom described sounds terrible.

Example Two

In another test item, test takers were presented the following jumbled words:

more, hormones, than, influence, adults, do

They were unable to identify the syntactical structure of the key without referring to hints either and thus made sentences mostly in random word orders. Some of their first attempts were:

More hormones than adults do influence

or

More adults do influence than hormones

and

Hormones do influence more than adults,

Once given the first hint, "*The sentence begins with 'Hormones,'*" test takers understood the sentence structure and made a meaningful sentence as follows:

Hormones do more than influence adults.

Here we should say it is not that the subjects know the sentence structure very well at the first time but that the subjects are believed to be better able to infer, with the hint(s) given, that the key of the JW test item must be a sentence in a complex structure. The examples, in some ways, justify the cognitive hypothesis: human knowledge is stored by means of propositional networks in the brain. The retrieval of the knowledge from the brain is achieved through the spread of activation. Thus, well-organized knowledge structure regarding attribute clause is activated more quickly with the help of hints and otherwise, more slowly. In case the retrieval is baffled, the test takers appear to lack the relevant grammar knowledge being tested. Hence, the provision of cues turns out to be an important way of diagnosing test takers' real language ability. The first attempts made by test takers in the examples given above also demonstrate that JW items do not allow for test takers' blind guessing.

RESULTS

The results of data analysis which examined item responses in addition to the latent factors underlying test performance are presented in this section.

Response Analysis

The present study obtained two ability curves, one indicating the curve based on response types and the other on both the response type and response time. As shown in Figure 1, response type and time values are highly correlated. The higher ability levels indicate response types A, B and E as these test takers spent less time in managing to get a correct arrangement of the jumbled words, while the low ability levels are those of type K, L and N. A further analysis of the responses showed that response time turned out to be a significant variable which was capable of distinguishing ability levels of the same response type.

Latent Factor Analysis

To verify the theoretical formalization described above, the present researcher applied Structural Equation Modeling (SEM) using EQS6.1 to investigate the concept of matching CCT traits with the expected model. Since the first application of SEM approach to language testing in 1981 (Bachman & Palmer, 1981), SEM has been used in a wide range of studies (e.g. Kunnan, 1995; Bae & Bachman, 1998). However, no studies in China have ever used a latent factor approach to address such fundamental issues about language abilities. Based on the above discussion, three latent factors have been specified which are believed in one way or another to influence the test taker's language ability. These three factors are (1) the test taker's mode of performance, (2) the test taker's condition and (3) the test item difficulty. According to SEM, these three latent factors are formed as three measurement models each containing four or five measured variables. Figure 2 shows the measurement model for test taker's mode of performance. The measured variables, V1 to V4, in the squared forms, indicate ST (Solution Time), HA (Hint-adopted), GG (Guessing) and CT (Cheating). These loaded on the latent variable Test Taker's Mode, i.e., the test taker's mode, showing how the test taker is coping with JW test items. Each single arrowed line expresses one variable affecting the other directly while each arrowed line pointing from E1, E2, and so on to the squared box indicates the un-interpretable parts of latent variables and can instead be understood as a kind of possible errors (Bentler,. & Wu, 2002; Bentler, 2006; Byrne, 1994; Jöreskog, 1970,1977; Bachman, 1998; Kunnan 1998,1999; Purpura,1998; Rob, 2005).

Figure 3 shows the measurement model for the test taker's condition containing the measured variables, V5 to V9, indicating TR (Test Readiness), TF (Test Familiarity), ID (Individual Difference), TRF (Test Room Familiarity), CBF (Computer-based Familiarity), TIF (Test Item Familiarity). These loaded on the latent variable Test Taker's Condition, or F2.

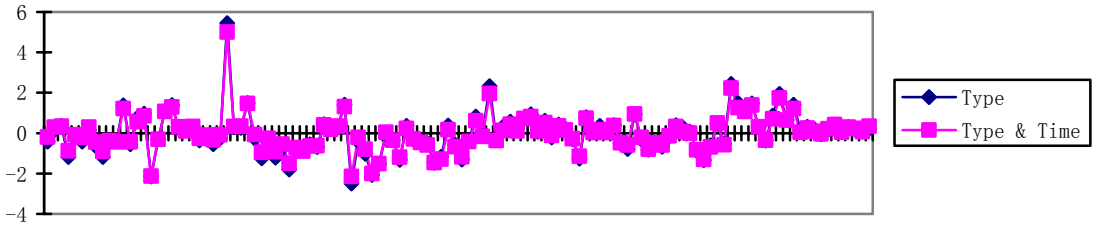


Figure 1. Ability curve based on the response type and response time (N=120).

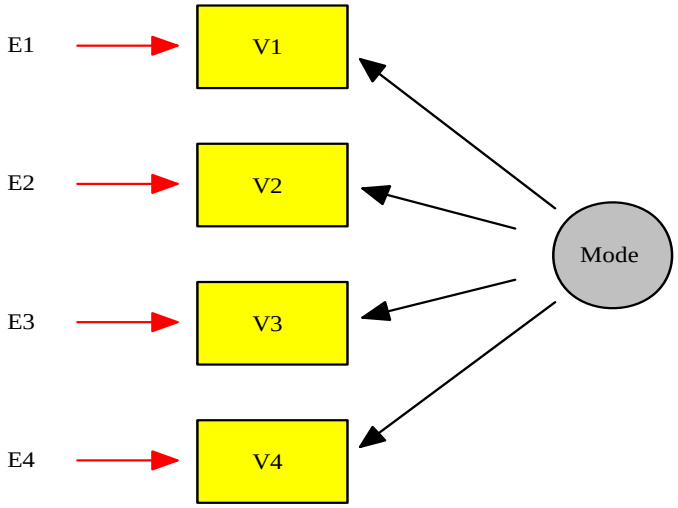


Figure 2. Measurement Model for Test Taker's Mode

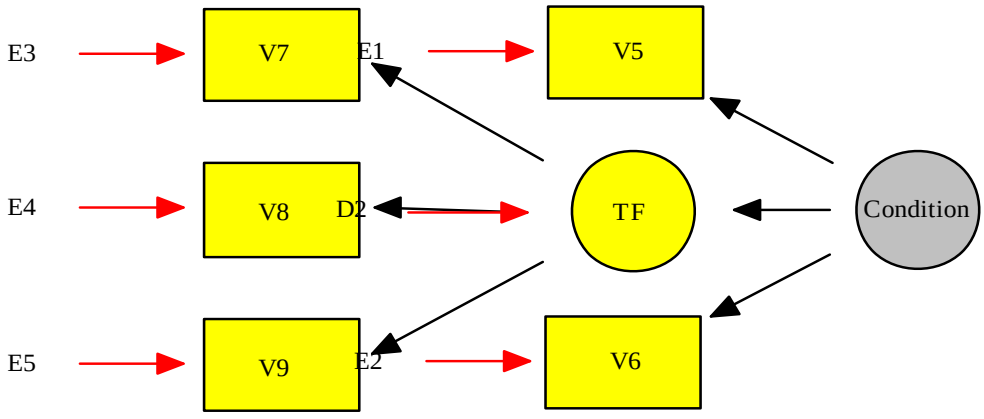


Figure 3. Measurement Model for Test Taker's Condition

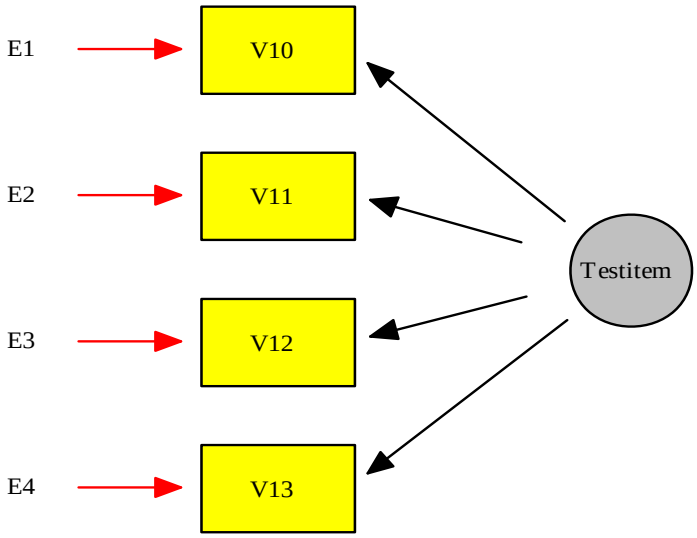


Figure 4. Measurement Model for JW Test Item Difficulty

Figure 4 presents the measured variables, V10 to V13, indicating WN (Word Number), SS (Sentence Structure), VOC (Vocabulary) and BG (Background Knowledge), which loaded on the latent variable JW Item Difficulty, or F3.

According to SEM principles, these measurement models are formulated in the confirmatory mode and are based on prior experimental results conducted during the researcher's doctoral studies. Parts of the data are the raw data collected from the test takers of PRETCO[v] administered from 2002-2005 in Guangdong Province, PR China. Figure 5 presents the second-order model[vi] using Test Taker's Mode, Test Taker's Condition and Test Item Difficulty linking both the independent and dependent variables and their associated measured variables and errors. According to SEM, such a model is based on the hypothesis that these three latent variables are structured as illustrated to represent the construct of language ability measured by this test.

As a second-order model, parameter estimates for measured variables and correlations among the latent variables are all calculated with EQS6.1. Figure 5 shows the output containing the goodness-of-fit statistics.[vii] The comparative fit index (CFI)[viii] = .931 indicates that the model is reasonably acceptable. However, as it is .931 rather than .95 or above, we may presume that there exist some other factor(s) that influence the measured language ability. Figure 6 shows some minor inappropriateness regarding the z-scores of variables like Test Room, Computer and Individual Difference as shown in Figure 6.

Figure 7 shows measurement equations with standard errors and test statistics after 198 cycles. As shown in the figure, both guessing and solution time are significant. But each of these z-scores for TRF (Test Room Familiarity), CBF (Computer-based Familiarity) and ID (Individual Difference) turns out to be negative, much smaller than the abstract

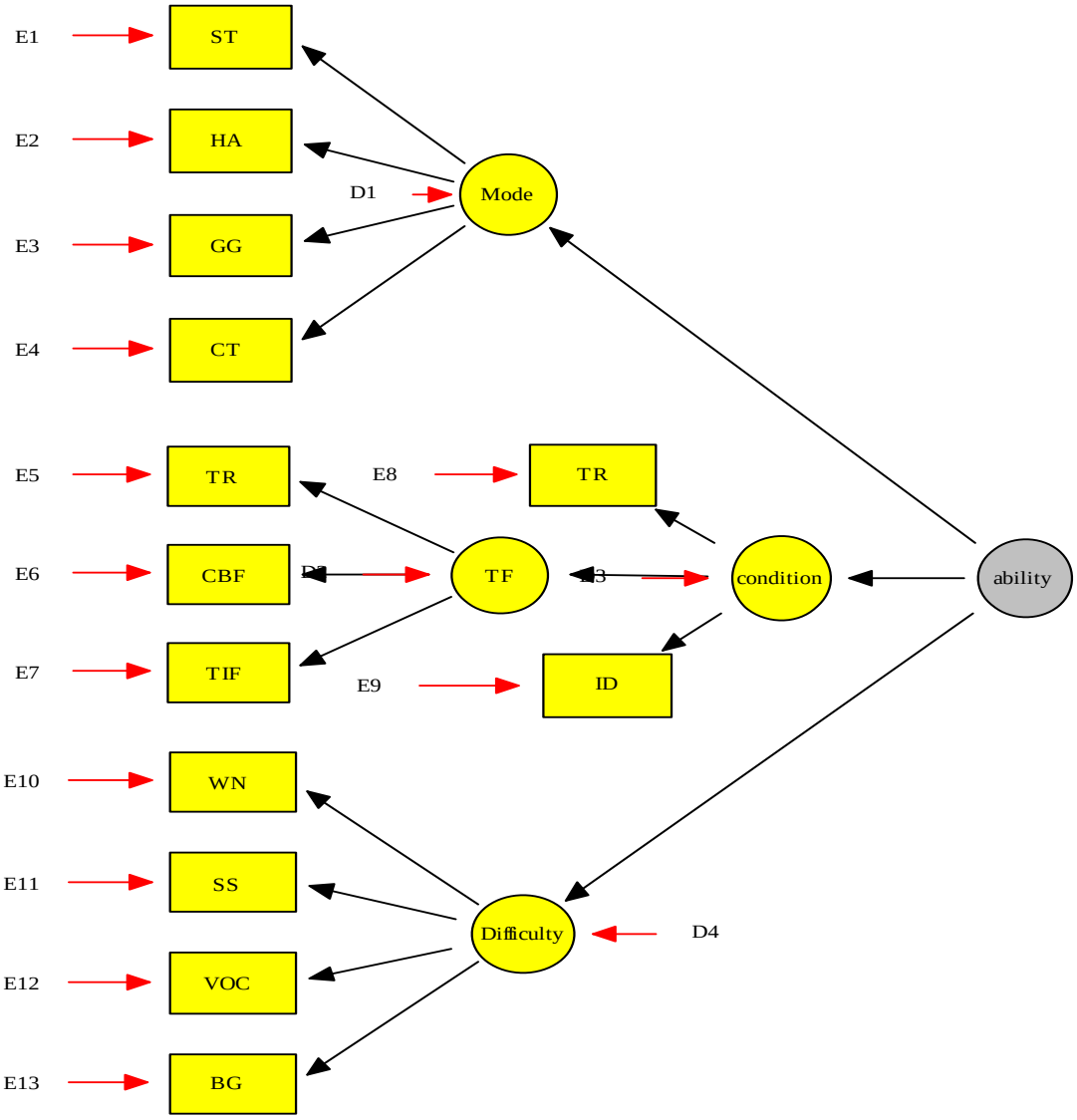


Figure 5. The Second-order Model for Test Taker's Mode, Test Taker's Condition and JW Test Item Difficulty

chi-square = 67.712 (df = 51)
probability value for the chi-square statistic = .05859
the normal theory rls chi-square for this ml solution = 63.756.
bentler-bonett normed fit index = .780
bentler-bonett non-normed fit index = .910
comparative fit index (cfi) = .931
root mean-square error of approximation (rmsea) = .052
90% confidence interval of rmsea (.000, .083)

Figure 6. Goodness of fit indices for structural model 1 (N=120)

value, 1.96, of 95% significance, suggesting that the measurement model for the test taker's condition (diagrammed previously in Figure 3) was not reasonably designed or that the data collection from questionnaires was problematic, or both. To be more exact, this can also be interpreted as an indication that the three potential construct-irrelevant variables of examinees' familiarity to the test room, examinees' familiarity to computers, and examinees' individual condition did not influence the language ability as measured.

Based on these results, the model was revised into a more parsimonious structural model as shown in Figure 8. The new model includes a bidirectional relationship between the latent variable Test Taker's Mode and the latent variable JW Item Difficulty. In this model the two correlated latent variables are each associated with four measured variables. This

GUESS = V2	=	.153*F1	+	1.0000 E2
		.022		
		7.093@		
TIME= V3	=	112.327*F1	+	1.000 E3
		10.230		
		10.980@		
CLASSRM=V6	=	-.760*F2	+	1.000 E6
		1.191		
		-.638		
COMPUTER=V7	=	-.103*F2	+	1.000 E7
		.188		
		-.546		
INDIVDAL=V9	=	-.371*F3	+	1.000 E9
		.303		
		-1.223		

Statistics significant at the 5% level are marked with @.

Figure 7. Measurement equations with standard errors and test statistics (Iteration = 198)

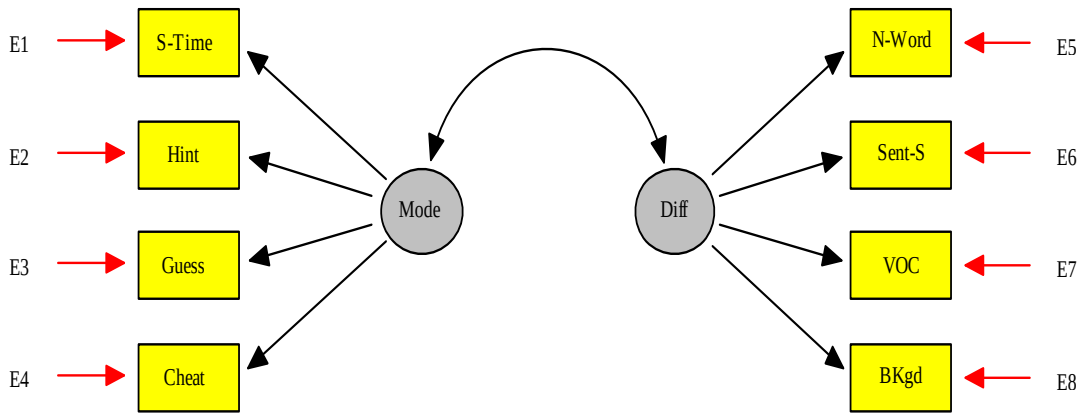


Figure 8. The revised structural model of CCT

model will need to be tested in subsequent research. Hopefully such research using SEM building under the guidance of cognitive science will help to test the CCT model and ultimately lead to significant diagnosis and estimation of language ability.

CONCLUSIONS

With the advent of the World Wide Web and the growth of the Internet, there is an increasing interest in expanding the availability of psychometric assessment services via the Internet. The traditional CAT provides some innovations over traditional linear testing that can be used for this purpose. However, a more significant expansion of assessment possibilities rests on the application of more sophisticated testing theory. The key aspects of assessment have been articulated as an "Assessment Triangle" consisting of cognition, observation and interpretation. The idea of Assessment Triangle was elaborated in the executive summary of "Knowing What Students Know: The Science and Design of Educational Assessment" compiled by The National Research Council (2001): "a model of cognition and learning, or a description of how people represent knowledge and develop competence in a subject domain, is a cornerstone of the assessment development enterprise. Unfortunately, the model of learning is not made explicit in many assessment development efforts, is not empirically derived, and/or is impoverished relative to what it could be" (p. 176).

This paper described an effort to move beyond an inexplicit assessment model to one that takes into account the three points of the triangle. From the perspective of cognitive science, the JW test item and its cognitive basis were elaborated; points of contrast between current CAT practice and CCT designs were discussed, and a pilot study of examinees' performance on such a test was conducted. It is believed that once CCT can be put into use, it will contribute to the evolution of practice in computer-based language testing. The present paper contributes to this evolution through SEM–based research to support CRT. One thing worth mentioning is that, although SEM approach to language testing via EQS has found a wider application and computer software like PARSCALE has been used for quite some time internationally, they have rarely been utilized in language assessment before this study. Hopefully, the research presented in this paper can be held as a good starting point for further investigation of diagnostic language assessment via computerized cognitive testing.

REFERENCES

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 60,* 451-474

Anderson, J. R. (1976). *Language, memory and thought.* Hillsdale NJ: Lawrence Erlbaum Associates

Anderson, J. R. (1983). *The Architecture of cognition*. Cambridge, MA: Harvard

University Press.

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: W.H. Freeman and Company.

Anderson, J. R., & Gluck, K. (2001). What role do cognitive architectures play in intelligent tutoring system? In S. M. Carver & D. Klahr (Eds.), *Cognition & instruction: Twenty-five years of progress* (pp. 227-261). Mahwah, NJ: Lawrence Erlbaum Associates.

Bachman, L. F. & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language learning*, *31*, 67-86.

Bachman, L. F. (1998). Modern language testing at the turn of the century: assuring that what we count counts. *Newsletter of the American Association for Applied Linguistics*, *21*(2), 11-13.

Bachman, L. F. (2006). Assessment Use Argument (AUA). Report presented at International Conference for English Teaching, Shantou University, Guangdong Province, PRC.

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing*, *15*(3), 380-414.

Bentler, P. M. & Wu, E.J.C. (2002). *EQS6 for Windows user's guide*. Encino,CA: Multivariate Software, Inc.

Bentler, P. M. (2006). *EQS6 structural equations program manual*. Encino,CA: Multivariate Software, Inc.

Binet, A. (1909). *Les idees morderne sure les enfants.* Paris: Ernest Flammarion

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Atkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443-469.

Bunderson, C. V., Inonye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), Educational Measurment (3rd ed.) (pp. 367-408). New York: American Council on Education/MacMillan Publishing Company.

Byrne, B. M. (1994). *Structural Equation Modeling with EQS and EQS Windows.* Thousand Oaks, CA: Sage.

Dillon, R. F. (1985). Predicting academic achievement with models based on eye movement data. *Journal of Psychoeducational Assessment, 3*, 157-165.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn, (Ed.) *Educational measurement* (3rd ed.) (pp. 147-200). New York: American Council on Education/MacMillan Publishing Company.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp.141-144). Washington DC: American Psychological Association.

Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. In Arthur S. Goldberger & O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp.85-112). New York/London: Seminar Press.

Jöreskog, K. G. (1977). Structural Equation Models in the Social Sciences: Specification estimation and testing. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 265-287). Amsterdam: North Holland.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 90,* 122-149.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*, 241-261.

Klahr, D., & Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. *Cognitive Psychology, 13*, 113-148.

Kunnan, A. J. (1995). *Test takers characteristics and test performance: a structural modeling approach*. Cambridge: Cambridge University Press.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*(3), 295-332.

Kunnan, A. J. (1999). Recent Developments in Language Testing. *Annual Review of Applied Linguistics, 19*, 235-253.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed), *Computer-assisted instruction, testing and guidance* (pp. 139-183). New York: Harper & Row.

Mislevy, R. J., & Verhelst, N. (1987). Modeling item responses when different subjects

employ different solution strategies. Technical Report RR-87-47-ONR, Educational Testing Service, Princeton, NJ.

National Research Council. (2001). *Knowing what students know*. Washington DC: National Academy Press.

Parshall, C. G., & Kromrey, J. D. (1993). Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect. Paper presented at the annual meeting of the American Education Research Association, Atlanta.

Purpura, J. (1996). Investigating the relationships between selected cognitive characteristics of test takers and performance on language tests. Unpublished doctoral dissertation, University of California, Los Angeles.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high-and-low-ability test takers; a structural equation modeling approach. *Language Testing, 15*(3), 333-379.

Reckase, M. D. (1973). An interactive computer program for tailored testing based on the one-parameter logistic model. Paper presented at the National Conference on the Use of On-line Computers in Psychology, St. Louis Mo.

Reese, C. (1992). Development of a computer-based test for the GRE general test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Revuelta, J., & Ponsada, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Education Research.

Schoonen, R. (2005). Generalizability of Writing Scores: An Application of Structural Equation Modeling. *Language Testing, 22*(1), 1-30.

Smith, R. M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement. 24*(23), 217-231.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn, (Ed.), *Educational measurement* (3rd ed.) (pp. 263-331). New York: American Council on Education/MacMillan Publishing Company.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.) *Computerized adaptive testing: A primer* (pp.65-102). Hillsdale, NJ: Erlbaum.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). Future challenges. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp.233-286). Hillsdale, NJ: Erlbaum.

Weiss, D. J., & Betz, N. E. (1973). Ability measurement: Conventional or adaptive? (Research Report 73-1), Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, (NTIS No. AD757788).

Weiss, D. J. (1974). Strategies of adaptive ability measurement (Research Report 74-5), Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, (pp.104-930).

Zhang, Q. (1993). Computerized cognitive testing: Theory, method and practice. Unpublished doctoral dissertation, Guangzhou Institute of Foreign Languages, Guangzhou, RPC.

Zhang, Q. (2002a). Computerized cognitive testing: A general introduction. Presentation made at ETS, Princeton, NJ.

Zhang, Q. (2002b). BILOG and Parscale: Different but Alike. Paper presented at Language Testing and Teaching, International Conference, Shanghai Jiaotong University.

APPENDIX A

TEN JUMBLED WORD TEST ITEM USED FOR CCT

Jumbled Word Test Item	Key to JW Test Item
hinders, too, calcium, growth, children's, much Hint 1: Begin with 'Too'; Hint 2: The word 'hinders' used as verb; Hint 3: This is a simple sentence structure.	Too much calcium hinders children's growth.
biologists, cultivated, oysters, to, spawn, induce Hint 1: Begin with 'Biologists' Hint 2: The word 'induce' used as verb; Hint 3: This is a simple sentence structure	Biologists induce cultivated oysters to spawn.
terrible, Tom, described, the, service, sounds, that Hint 1: Begin with 'The' Hint 2: The word 'that' used as relative pronoun Hint 3: This sentence contains an imbedded attribute clause.	The service that Tom described sounds terrible.
more, hormones, than, influence, adults, do, Hint 1: Begin with 'Hormones' Hint 2: The word 'do' used as verb; Hint 3: 'more than' used as collocation.	Hormones do more than influence adults.
Awhile, glaciers, float, and melt, about Hint 1: Begin with 'Glaciers' Hint 2: The word 'float' used as verb; Hint 3: This is a simple sentence structure.	Glaciers float about awhile and melt.
what, is, their most computers, matters Hint 1: Begin with 'What'; Hint 2: the word 'is' used a verb; Hint 3: This sentence contains a subject clause.	What matters most is their computers.
they, do, left, make, with, margarine Hint 1:Begin with 'They'; Hint 2: The word 'left' used a post-modifier; Hint 3: This is a simple sentence structure.	They make do with margarine left.
complain, beaver, dams, fishing, enthusiasts, about Hint 1: Begin with 'Fishing'; Hint 2:The word 'complain used as verb;' Hint 3: This is a simple sentence structure.	Fishing enthusiasts complain about beaver dams.
would, further, delay, us, greater, cause, losses Hint 1: Begin with 'Further'; Hint 2: The word 'cause used as verb;' Hint 3: This is a simple sentence structure.	Further delay would cause us greater losses.
A, reelection, win, cartoon, helped, him Hint 1: Begin with 'A'; Hint 2: The word 'helped' used a verb; Hint 3: This is a simple sentence structure.	A cartoon helped him win reelection.

APPENDIX B

PARSCALE COMMAND FILE FOR PARTIAL CREDIT MODEL

```
CCTJW01.PSL      TOWARDS COGNITIVE RESPONSE THEORY (JUMBLED WORD DATA)
                  GENERALIZED PARTIAL CREDIT MODEL - EAP SCALE SCORES
>COMMENTS
    This example scores and calibrates the data of categorical response type with response time assuming the partial credit model with standard scoring function.
    To illustrate the situation where 10 jumbled word items are involved, each with 3 relevant hints provided. Totally, 16 categories for the response type are specified.
    The standard score function assumes 16 is the high category, so response modification is required in BLOCK1.
    Thus, for response to each item produced by a test taker, there are two records: response type and response time.
    As PARSCALE accepts ordinal data, the real-valued response time presented by test takers is converted into six categories coded: Native User, Near Native User, Good User, Modest User, Average User and Poor User.
    The items are analyzed in two subtests. The first subtest consists of 10 response types and the second, of 10 response time codes.
    The data file contains the test taker ID, followed by the 10 response type and time code

>FILES     DFNAME='CCTWJ03.DAT', SAVE;
>SAVE       SCORE='CCTWJ03.SCO', COMBINE='CCTWJ03.CMB';
>INPUT     NIDW=9, NTOTAL=20, NTEST=2, LENGTH = (10,10), COMBINE=2;
 (9A1, 1X, 20A1)
>TEST1     TNAME='TYPE', ITE = (1(1)10), NBLOCK=1, SLOPES=(1.0(0)10), THRESHOLDS=(0.0(0)10);
>BLOCK1    BNAME='BLK-TYPE', NIT=10, NCAT=15, ORIGINAL=(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O),
            MODIFIED=(15,14,13,12,11,10,9,8,7,6,5,4,3,2,1);
>CALIB     PARTIAL, LOGISTIC, NQPTS=31, CYCLE = 100, NEWTON=2, CRIT=0.001, SCALE=1.7, SPRIOR;
>SCORE     MLE, SMEAN=0.0, SSD=1.0, NAME='PCM_MLE', PFQ=5;
>TEST2     TNAME='TIME', ITE = (11(1)20), NBLOCK=1, SLOPES=(1.0(0)10), THRESHOLDS=(0.0(0)10);
>BLOCK2    BNAME='BLK-TIME', NIT=10, NCAT=6, ORIGINAL=(A,B,C,D,E,F),
            MODIFIED=(6,5,4,3,2,1);
>CALIB     PARTIAL, LOGISTIC, NQPTS=31, CYCLE = 100, NEWTON=2, CRIT=0.001, SCALE=1.7, SPRIOR;
>SCORE     MLE, SMEAN=0.0, SSD=1.0, NAME='PCM_MLE', PFQ=5;
>COMBINE NAME=STRAIGHT, WEIGHTS=(0.5,0.5);
>COMBINE NAME=STRAIGHT, WEIGHTS=(0.9,0.1);
```

[i] For detailed limitations of current assessment, interested readers may refer to pp.26-29 in *Know what students know: The science and design of educational assessment*. National Academy Press. Washington, DC. 2001.

[ii] For details about "Assessment Triangle", see p.2, *Know what students know: The science and design of educational assessment*. National Academy Press. Washington, DC. 2001.

[iii] For detail, see reaction-time studies ,p. 98. National Research Council. (2001). *Knowing what students know.* Washington DC: National Academy Press. USA

[iv] For detail, see reaction-time studies ,p. 99. National Research Council. (2001). Knowing What Students Know. Washington DC: National Academy Press. USA

[v] RETCO is abbreviated from Practical English Test for Colleges administered twice a year in technical and vocational institutes and colleges in China with the total number of candidates reaching over a million a time. The author has been the chief examiner of PRETCO at Guangdong Provincial level since 1998.

[vi] The relevant structural equation model is currently under moderation based on the information given from LM Test.

[vii] Briefly, goodness-of-fit yielded via EQS6.1 is actually referred to the kind of matching or approximation parameter regarding the observed data to the expected model after certain designated iterations (In our case, 300 cycles were set). It can be also understood as function of the data measuring the distance between the hypothesis and the data and the probability of obtaining data. The most common tests for goodness-of-fit are the chi-square test, Kolmogorov test, and Cramer-Smirnov-Von-Mises test. EQS6.1 uses chi-square test.

[viii] According to Kunnan (1998), generally, if any of these indices are above .90, the thumb is that there is recommendation from the indices that there is a model fit, pending examination of the Chi square statistic and model interpretability.

Part II

NLP Analysis in Language Assessment

Cotos, E. & Pendar, N. (2008). Automated diagnostic writing tests: Why? How? In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 65-81). Ames, IA: Iowa State University.

Automated Diagnostic Writing Tests:
Why? How?

Elena Cotos
Nick Pendar
Iowa State University

Diagnostic language assessment can greatly benefit from a collaborative union of computer-assisted language testing (CALT) and natural language processing (NLP). Currently, most CALT applications mainly allow for inferences about L2 proficiency based on learners' recognition and comprehension of linguistic input and hardly concern language production (Holland, Maisano, Alderks, & Martin, 1993). NLP is now at a stage where it can be used or adapted for diagnostic testing of learner production skills. This paper explores the viability of NLP techniques for the diagnosis of L2 writing by analyzing the state of the art in current diagnostic language testing, reviewing the existing automated scoring applications, and considering the NLP and statistical approaches that appear promising for automated diagnostic writing assessment for ESL learners.

INTRODUCTION

In language assessment, diagnostic language tests are defined as those that aim to identify learners' areas of strength and weakness (Alderson et al., 1995; Bachman & Palmer, 1996; Davies et al., 1999; Moussavi, 2002) in order to help improve learning. The strengths identified should point to the level a learner has reached, and the weaknesses detected should indicate areas for improvement. Alderson claims that diagnostic tests are the "closest to being central to learning" a second or foreign language (2005, p. 4). However, he also points out that diagnosis in second language testing lacks a clear theoretical basis, is under-investigated, and therefore, is underrepresented in the field. Despite the intuitive potential of diagnostic testing, the practical barriers to progress in this area include the need for a means of producing and storing detailed information about examinees' performance. In educational settings, such requirements seem to necessitate the use of technology.

In this paper, we argue that computer-assisted language testing (CALT)—and particularly diagnostic testing—would benefit from employing automated scoring systems such as those used in high-stakes standardized writing tests. We point out the advantages of the proposed automated writing tests and then emphasize key directions for moving forward on this research agenda. We begin by addressing questions in the design of such tests and the options for test items. Since Automated Essay Scoring (AES) systems evaluate constructed responses, we will then closely examine AES programs and the natural

language processing (NLP) approaches they employ, which appear to be particularly promising for automated diagnostic writing assessment. Finally, we will discuss issues in the validation of such tests. In conclusion, we call for future research on diagnostic assessment and for incremental collaboration among specialists in areas related to language learning.

ADVANTAGES OF AUTOMATED WRITING TESTS

Automated scoring would be a promising innovation for diagnostic writing assessment. Dikli (2006) emphasizes that automatic scoring systems can enhance practicality, helping overcome time and cost issues. Assessment of writing has traditionally implied design of prompts, creation of rubrics, training of raters, and scoring the responses by humans. Indisputably, automated scoring can reduce the need for some of these activities because once the scoring system is built it can automatically evaluate the qualities of examinees' performance (Williamson, Mislevy, & Bejar, 2006) by analyzing evidence that would allow for making inferences about strengths and weaknesses in learners writing ability. Moreover, if substantial information can be gained from such performance, the system's analyses of constructed responses could both describe learners' performance and place them in an appropriate level. This would make it possible to eliminate an initial placement procedure used in certain tests. In fact, because learners' written production can be analyzed in such great detail, one can argue that there would be no need for designing separate tests for individual skills such as grammar, vocabulary, etc.

With respect to reliability, essay grading is criticized for "perceived subjectivity of the grading process" (Valenti, Nitko, & Cucchiarelli, 2003, p. 319) because of the frequent variation in scores assigned by different raters. Automated evaluation could increase objectivity of assessment, providing consistency in scoring and feedback through greater precision of measures (Phillips, 2007). Also, the systems, if re-trained, would be able to re-score student answers should the evaluation rubric be redefined (Rudner & Gagne, 2001). Finally, automated diagnostic tests could have built-in validity checks to address possible biases (Page, 2003).

A third advantage is related to diagnostic assessment's provision of meaningful feedback, which Heift (2003) defines as a "response that provides a learning opportunity for students." (p. 533). The characteristics of feedback that is likely to prove meaningful to examinees are likely to be similar to those identified in research on second language learning. Table 1 lists types of feedback that show promise based on the studies indicated. If automated diagnostic testing resulted in such feedback returned to learners, it may be possible for them to take steps towards remediation and improvement. Diagnostic tests might also enhance learning opportunity by allowing learners to act upon the received feedback, re-submit their texts, and make gradual improvements. Moreover, because automatic scoring systems are generally trained on certain material, directed feedback could be linked to the training texts (Landauer, Laham, & Foltz, 2003) (which could be either model or learner texts), thus making diagnostic assessment interactive, tailored

Table 1. Feedback leading to better learning and research investigating its use

1. Explicit feedback (Caroll, 2001; Caroll & Swain, 1993; Ellis, 1994; Lyster, 1998, Muranoi, 2000)
2. Individual specific (Hyland, 1998)
3. Metalinguistic feedback (Rosa & Leow, 2004)
4. Negative cognitive feedback (Ellis, 1994; Long, 1996; Mitchell & Myles, 1998)
5. Intelligent feedback (Nagata, 1993, 1995)
6. Output-focused feedback (Nagata, 1998)
7. Detailed iterative feedback (Hyland & Hyland, 2006)
8. Feedback – accurate, short, one at a time (Van der Linden, 1993)

both to instruction and to individual learners. For examples of systems that have already implemented tools which produce feedback oriented toward instruction, interested readers can look into *Criterion*^{SM} by Educational Testing Service and *MY Access* by Vantage Learning.

Finally, as Xi (this volume) points out, automated evaluation would not be a mere application of new technologies; it would become an essential component of the validity argument for the use of automated diagnostic tests. Moreover, the focus on evidentiary reasoning would facilitate the development of automated diagnostic tests if we choose to follow the framework of Evidence-Centered Design, which "is an approach to constructing and implementing educational assessments in terms of evidentiary arguments" (Mislevy, Steinberg, Almond, & Lukas, 2006, p. 15). With these potentials of automated diagnostic writing assessment, it is worth examining how such tests can be designed.

THE DESIGN OF DIAGNOSTIC TESTS

Although "virtually any test has some potential of providing diagnostic information" (Bachman, 1990, p. 60), some guidelines exist for the design of diagnostic tests. According to Schonell and Schonell (1960), such tests should not impose time limits. Bejar (1984) distinguishes a diagnostic test from other types of assessment by the fact that a diagnostic test is self-referencing. In achievement and norm-referenced tests, for instance, referencing is typically with respect to a population, while "in a diagnostic test the student's performance is compared against his or her expected performance" (Bejar, 1984, p. 176). Furthermore, a diagnostic test should be oriented towards learning by providing students with explicit feedback to be acted upon in addition to displaying immediate results. It should generate a detailed analysis of learner responses, which should lead to remediation in instruction.

However, the central issue in test design is what should a diagnostic test evaluate to

reveal the learner's relevant strengths and weaknesses? How closely should diagnostic tests be aligned with a particular curriculum or materials? One approach to the design of diagnostic testing is to create the test specifications on the basis of content that is taught in the textbooks or CALL materials that they are intended to accompany. The feedback that students receive from such a test can refer students back to specific parts of the materials. Irrespective of the kind of instruction, it can be based on the content that has been or will be covered in the teaching process and become an essential part of individualized instruction or self-instruction. Unlike many other tests, its results should be qualitative or analytic rather than quantitative, and their interpretation should not be used for high-stakes decisions.

The other approach to the design of diagnostic tests is to base diagnostic information on theoretical perspectives on the development of second language proficiency. As Alderson (2005) puts it, "[w]ithout a theory of development, a theory, perhaps also, of failure, and an adequate understanding of what underlies normal development as well as what causes abnormal development or lack of development, adequate diagnosis is unlikely" (p. 25). A theory of language development is important in language testing for purposes of construct definition and level scale generation, and this is the central concern for researchers in second language acquisition (SLA).

In the absence of useful theoretical perspectives in second language acquisition, a number of developmental frameworks have been elaborated; e.g., ACTFL scales (American Council for Teaching of Foreign Languages, 1983), International Language Proficiency scales (Wylie & Ingram, 1995/1999), Canadian Benchmarks (Pawlikowska-Smith, 2000), and the Common European Framework of Reference (CEFR) (Council of Europe, 2001). A diagnostic test based on the CEFR provides an example of how test designers might use such frameworks for test design. DIALANG, a unique piloting effort to develop and implement computer-based diagnostic tests, was a European Union-funded project intended to provide diagnostic information about learners' reading, listening, writing, grammar, and vocabulary proficiency in 14 languages relying on CEFR. The test results were to be interpretable on the CEFR scale which was intended to be useful for students in many different situations.

The main aspects that are targeted by the writing section of DIALANG are *textual organization*, *appropriacy*, and *accuracy* in writing for communicative purposes such as providing information, arguing a point, or social interaction. For textual organization, learners are diagnosed based on how good they are at detecting coherence and cohesion markers; for appropriacy, based on how well they can set the tone and the level of formality in the text; and for accuracy, based on how they can cope with grammar and mechanics. For the latter, Alderson (2005) provides a somewhat detailed[i] frame of grammatical structures[ii] (See Table 2).

Assessment of writing proficiency would be incomplete without an analysis of learners' vocabulary. DIALANG incorporates separate vocabulary tests, which are targeted at learners' knowledge of the meanings of single words and word combinations.

Table 2. Morphological and syntactical categories

Morphology		**Syntax**	
Nouns	Inflection – cases Definite/indefinite – articles Proper/common	*Organization/ Realization of Parts of Speech*	Word order – statements, questions, exclamation agreement
Adjectives and Adverbs	Inflection Comparison	*Simple and Complex Clauses*	Coordination Subordination Deixis
Pronouns	Inflection Context	*Punctuation*	
Verbs	Inflection – person, tense, mood, active/passive voice		
Numerals	Inflection Context		

Specifically, knowledge of vocabulary is evaluated from several perspectives – word formation by affixation and compounding; semantic ties between synonyms, antonyms, hyponyms, polysemantic words, etc.; word meanings including denotation, connotation, semantic fields; and word combinations such as idioms and collocations.

Although DIALANG is brought into this discussion only as an example of how specific areas of writing ability can be defined, its construct definitions cover the most essential writing subskills, and, therefore, appear to also be appropriate for automated diagnosis of constructed responses further considered in the paper. However, modifications can certainly be made depending on the specificity with which test-developers intend to approach the diagnostic task.

Regardless of whether the test design relies on course materials or on a general framework, the implementation of the test requires a reliable means of gathering, evaluating, and storing relevant aspects of learners' performance. These operational issues are what we are concerned with in this paper. Obtaining detailed profiles of learner written performance across various components of the construct for diagnosing writing ability appears to be possible if NLP-based automated scoring is employed by CALT.

COMPUTER-BASED DIAGNOSTIC WRITING TEST ITEMS

Samples of examinees' performance can be obtained using a variety of test items or tasks. The requirements of the automated scoring procedure depend in part on the degree of constraint placed on the examinee's response. Scalise and Bernard (2006) provide a comprehensive taxonomy for electronic assessment questions and tasks that include multiple choice, selection/identification, reordering/rearranging, substitution/correction,

completion, construction, and presentation/portfolio. Existing diagnostic tests, however, still follow the constrained approach, in which components of the construct are assessed "indirectly through traditional objectively assessable techniques like multiple choice" (Alderson, 2005, p. 155). Indeed, our example, DIALANG, consists of such item formats as multiple choice, drop-down menus, text-entry, and short-answer questions. While these item types are not without merit, they are often criticized for lacking what some people call face validity, credibility in the eyes of test users as measures of the intended construct (Williamson et al., 2006, p. 4).

This criticism is particularly apt in the testing of second language writing because selected response tasks fail to draw upon the productive abilities of interest, and therefore any relationship between test performance and the abilities of interest as very indirect. Perspectives on second language acquisition such as interactionism, socioculturalism, and functionalism, attribute a central role to *output*, considering it to be the real evidence that learners acquired certain linguistic phenomena (Ortega, 2007). Selected response measurement can only assess learners' ability to comprehend and choose among options in the input, which may be rather suitable for obtaining information about learners' receptive language skills such as reading and listening. However, indirect test items are not capable of leading to accurate inferences about learners' writing and speaking because they do not obtain information on how well learners integrate the input and how well they can produce output in the target language. In order to provide accurate diagnosis of learners' strengths and weaknesses of productive skills, we need to elicit more than recognition; we need to evaluate learners' output, or production.

In view of the need to gather samples of examinees' language production, diagnostic writing assessments need to expand on currently used techniques by adding constructed response tasks (Bennett & Ward, 1993). Williamson et al. (2006) emphasize the educational value of such items. Based on their analysis of the research in this area, they argue that constructed responses are beneficial because they

- "are believed to be more capable of capturing evidence about cognitive processes"
- "provide better evidence of the intended outcomes of educational interventions"
- "offer better evidence for measuring change on both a quantitative […] and a qualitative level […]," and
- "permit the opportunity to examine the strategies that examinees use to arrive at their solution, which lends itself to collecting and providing diagnostic information" (p. 4).

These points are made with respect to testing of a variety of content; however, in language assessment, the issue is even more straightforward: if learners' strengths and weaknesses in writing ability are to be detected, they need to write! Only by observing

their extended performance, i.e., how well they can produce texts that are comprehensible, intelligibly organized, register appropriate, correctly punctuated, etc., can we judge their writing proficiency. Moreover, constructed responses based on an adequate task can exhibit various contexts created by learners as well as multiple examples of grammatical structures in use, allowing us to obtain a detailed analysis of their command of grammar. As for vocabulary, these test items would bring diagnosis to the next level by revealing learners' ability to operate with words in order to create comprehensive contexts.

Constructed responses are also advantageous from the viewpoint of practicality. Designing selected response computer-based diagnostic tests as well as any other types of tests requires considerable effort, especially when it comes to test items. It is very laborious to develop specifications, create a good size pool of items, and pilot the items in order to select the ones that are reliable. In contrast, diagnostic tests based on constructed responses would be more time and cost-efficient in that the test developers would only develop effective prompts. These could be, for instance, essay prompts similar to the ones used in TOEFL, or open-ended questions requiring description, comparison, hypothesizing, etc. Further, Alderson (2005) admits that DIALANG designers "recognized the impossibility of developing specifications for each CEFR-related level separately" (p. 192). This may be less of a problem for constructed response tasks due to the prompts. When the same prompt is used by learners of different levels of proficiency, it is their performance that will differ, resulting in different diagnoses as well.

AUTOMATED SCORING SYSTEMS

The theory and practices of automated scoring are not covered by a single phrase. They are referred to as computerized essay scoring, computer essay grading, computer-assisted writing assessment, or machine scoring of essays, and existing systems go by terms such as AEG (Automated Essay Grading), AES (Automated Essay Scoring), and AWE (Automated Writing Evaluation). Despite the numerous terms, these practices are based on "the ability of computer technology to evaluate and score written prose" (Shermis & Burstein, 2003, p. xiii). The earlier computerized evaluation systems focused on essays, which can be seen in their names, but more recent innovations have expanded the concept of written prose and now include free text or short response answers.

Dikli (2006), Phillips (2007), and Valenti et al. (2003) provide a comprehensive view of existing AES systems, describing their general structure and performance abilities and discussing issues related to their use in testing as well as in the classroom. Here, we will briefly review the most widely used systems in order to further show that their functionality can be extrapolated to diagnostic assessment.

One of the pioneering projects in the area of automated scoring was *Project Essay Grade* (PEG), which was developed in 1966 "to predict the scores that a number of competent human raters would assign to a group of similar essays" (Page, 2003, p. 47). It mainly

relies on an analysis of surface linguistic features of the text and is designed based on the concepts of *trins* and *proxes*. Trins represent intrinsic variables such as grammar (e.g., parts of speech and sentence structure), fluency (e.g., essay length), diction (e.g., variation in word length), etc., while proxes are the approximations or correlations of those variables, referring to actual counts in student texts. Focusing on writing quality, and based on the assumption that quality is displayed by the proxes, PEG relies on a statistical approach to generate a score. Recently, PEG has gone through significant modifications, e.g., dictionaries and parsers were acquired, classification schemes were added and tested, and a web-based interface has been developed.

In the late 1990s, the Pearson Knowledge Analysis Technologies produced the *Intelligent Essay Assessor* (IEA) – a set of software tools developed primarily for scoring content related features of expository essays. In order to measure the overall quality of an essay, IEA needs to be trained on a collection of domain-representative texts. It is claimed to be suitable for analysis and rating of essays on topics related to science, social studies, history, business, etc. However, it also provides quick customized tutorial feedback on the form related aspects of grammar, style, and mechanics (Landauer, Laham, & Foltz, 2003). Additionally, it has the ability to detect plagiarism and deviant essays. IEA is based on a text analysis method, Latent Semantic Analysis (LSA), and, to a lesser extent, on a number of Natural Language Processing (NLP) methods. This allows the system to score both the quality of conceptual content of traditional essays and of creative narratives (Landauer et al., 2003) as well as the quality of writing.

The *Electronic Rater* (E-Rater) is a product from the Educational Testing Service that has been used for operational scoring of the Graduate Management Admissions Test (GMAT) Analytical Writing Assessment since 1999. E-Rater produces a holistic score after evaluating the essay's organization, sentence structure, and content. Burstein (2003) explains that it accomplishes this with the help of a combination of statistical and NLP techniques, which allow for analyses of content and style. For its model building, E-Rater uses a corpus-based approach which differs from a theoretical approach in which features are hypothesized based on characteristics expected to be found in the essays. The e-rater corpus contains unedited first-draft essays. Outputs for model building and scoring are provided by several independent modules. The syntactic module is based on a parser that captures syntactic complexity; the discourse module analyzes the discourse-based relationship and organization with the help of cue words, terms, and syntactic structures; and the topical analysis module identifies the vocabulary use and topical content.

In addition to E-Rater, *IntelliMetric*, a product of Vantage Learning, has been employed for the rating of the Analytical Writing Assessment section of the GMAT since 2006. It is the first automated scoring system that was developed on the basis of artificial intelligence (AI) blended with NLP and statistical technologies. IntelliMetric is "a learning engine that internalizes the characteristics of the score scale [derived from a trained set of scored responses] through an iterative learning process," creating a "unique solution for each stimulus or prompt" (Elliot, 2003, p. 71). To attain a final score, more than 300 semantic, syntactic, and discourse level features are analyzed by this system.

They can be categorized into five groups: focus and unity (i.e., cohesiveness and consistency in purpose and main idea), development and elaboration (i.e., content through vocabulary use and conceptual support), organization and structure (i.e., logical development, transitional flow, relationship among parts of the response), sentence structure (i.e., syntactic complexity and variety), and mechanics and conventions (i.e., punctuation, sentence completeness, spelling, capitalization, etc.). Apart from the scoring ability, IntelliMetric's modes allow for student revision and editing as well as for diagnostic feedback on rhetorical, analytical, and sentence-level dimensions.

The *Bayesian Essay Test Scoring System* (BETSY), funded by the Department of Education and developed at the University of Maryland, was also designed for automated scoring. BETSY relies on a statistical technique based on a text classification approach that, as Valenti et al. (2003) claim, may combine the best features of PEG, LSA, and E-Rater. A large set of essay features are analyzed, among which are content-related features (e.g., specific words and phrases, frequency of content words) and form-related features (e.g., number of words, number of certain parts of speech, sentence length, and number of punctuation marks). Rudner and Liang (2002) assert that this system can also be used in the case of short essays, applied to various content areas, employed to provide a classification on multiple skills, and allow for obtaining diagnostic feedback in addition to scoring.

The *Automark* software system was developed in the UK in 1999 as an effort to design robust computerized marking of responses to open-ended prompts. The system utilizes NLP techniques "to perform an intelligent search of free-text responses for predefined computerized mark scheme answers" (Mitchell, Russel, Broomhead, & Aldridge, 2002, pp. 235-236). Automark analyzes the specific content of the responses, employing a mark scheme that indicates acceptable and unacceptable answers for each question. The scoring process is carried out by a number of modules: syntactic preprocessing, sentence analysis, pattern matching, and feedback. The latter is provided as a mark, but more specific feedback is also possible (Valenti, 2003). What makes it similar to human raters is the fact that, while assessing style and content, it can ignore errors in spelling, typing, syntax, and semantics that do not interfere with comprehension. All of these systems show great promise for automatic essay scoring, but they do so by taking a variety of approaches to analysis.

TECHNIQUES AND CONSTRUCTS

To analyze the constructed input and to produce scores and feedback, each of the systems described above uses one or a combination of statistical, natural language processing and artificial intelligence approaches. Moreover, each system targets somewhat different constructs as the aim of measurement procedures.

Statistical approaches to essay evaluation tackle the problem from the perspective of identifying sequences of textual features that, with some degree of probability, are likely to appear in texts of a known level of quality. As a consequence, a corpus of texts of

known quality is required to serve in an initial training phase for parameter estimation. The actual statistical analyses can be conducted in a number of different ways. For example, E-Rater employs "simple keyword analysis," which looks for coincident keywords between the student essay and the scored one. PEG relies on "surface linguistic features analysis" that finds the features to be measured and uses them as independent variables in a linear regression to yield the score. IEA, in turn, is underpinned by "latent semantic analysis (LSA)," a complex statistical technique developed for information retrieval and document indexation (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). LSA finds repeated patterns in the student response and the reference text to extract the conceptual similarity between them. Finally, BETSY is based on "text categorization" techniques, which can consist of several score categories, associate the student response with one of them, and assign the score accordingly.

Natural language processing techniques apply methods from computational linguistics for the analysis of natural language (Burstein, 2003). Based on linguistic rules that define well-formed, and in some cases erroneous, syntactic constructions, NLP techniques include syntactic parsers that evaluate the linguistic structure of a text. More recently, rhetorical parsers have also been developed to analyze the discourse structure of texts based on rules. Combining NLP with statistical techniques can result in systems that produce deep-level parsing and semantic analysis, therefore gathering more accurate

Table 3. Techniques used in automated scoring systems.

System	**Constructs**	**Technique**
PEG (Page, 2003)	Grammar, fluency, diction	Statistical (measurement of surface linguistic features)
IEA (Landauer et al., 2003)	Content Grammar, style, mechanics Plagiarism and deviance	Statistical (Latent Semantic Analysis)
E-Rater (Burstein, 2003)	Topical content Rhetorical structure Syntactic complexity	Statistical (e.g., vector analyses) Natural Language Processing (NLP) (e.g., part-of-speech taggers)
BETSY (Rudner and Liang, 2002)	Content Grammar, style, mechanics	Statistical (Bayesian text classification)
IntelliMetric (Elliot, 2003)	Focus / unity Development / elaboration Organization / structure Sentence structure Mechanics / conventions	Artificial Intelligence (AI) Natural Language Processing (NLP) Statistical
Automark (Mitchell et al., 2002)	Content Grammar, style, mechanics	Natural Language Processing (NLP)

information about the student's response and potentially providing a more accurate assessment. Among the current scoring systems, E-Rater, Automark,[iii] and IntelliMetric successfully employ NLP. IntelliMetric, in addition to NLP, exploits artificial intelligence techniques.

Artificial Intelligence techniques refer to computer programs that encode some procedures for reasoning and decision making about data that the program is provided. In the case of automatic essay analysis the reasoning and the decision-making that the program is to do is the assignment of scores to an essay, and the data are the essays that the program is to rate. Dikli (2006) claims that the IntelliMetric system is "modeled on the human brain." It is based on a "neurosynthetic approach […] used to duplicate the mental processes employed by the human expert raters" (p. 17). Apparently, the underlying scoring mechanism in ItelliMetric is a neural network (see Baum, 2004).

The approaches used in these systems are summarized in Table 3, which also includes the writing constructs that the various systems aim to measure. The constructs include aspects of writing quality that most writing teacher would recognize as important aspects of writing such as grammar, style, mechanics, plagiarism, topical content, and rhetorical structure. Despite the importance of these aspects of writing, human ratings of these areas are notoriously time-consuming and unreliable. Automated scoring systems can, in principle, assess these, plus other construct components (see Table 3); moreover, they can do that with precision and objectivity which may improve the assessment of writing for diagnosis.

In view of the functionality of existing systems, the potential of scoring systems for diagnostic assessment of ESL writing is undeniably apparent. However, as Xi (this volume) explains, an essential aspect of the research in this area are studies that demonstrate the validity of the systems for making the intended inferences about examinees' abilities.

VALIDATION RESEARCH

Recent empirical work provides evidence that E-Rater, IEA, PEG, IntelliMetric, Automark, and BETSY are valid and reliable (Burstein, 2003; Elliot, 2003; Keith, 2003; Landauer et al., 2003; Mitchell et al., 2002; Page, 2003; Valenti et al., 2003). The main method employed for system validation is single essay agreement results with human ratings. Summarizing research results, Dikli (2006) concludes that correlations and agreement rates between the system and human assessors are typically high. Experiments on PEG obtained a multi-regression correlation of 87%. E-Rater has scored essays with agreement rates between human raters and the system consistently above 97%. BETSY achieved an accuracy of over 80%. Automark's correlation ranged between 93% and 96%. IEA yielded a percentage for an adjacent[iv] agreement with human graders between 85% and 91%. IntelliMetric also reached high adjacent agreement (98%), and the correlation for essays not written in English attained 84%.

These results showing correlations between human and computer ratings would, of course, serve as only one part of a larger validity argument for the intended interpretations and uses of the systems. Moreover, the validity arguments to be made concerning each of these systems are for inferences about the writing of native speakers of English. While there is no doubt that their ability to analyze free production would be extremely valuable in assessing non-native speaker responses, there might be questions as to whether such systems can be as reliable in the case of ESL/EFL. Indeed, computerized assessment of constructed responses produced by non-native speakers, especially at low levels of proficiency, is prone to face barriers in dealing with ill-formed utterances.

Research in this area is only beginning; however, recent implementations and insights seem to be encouraging. For instance, in practical terms, Educational Testing Service has been successfully employing E-Rater to evaluate ESL/EFL performance on the TOEFL exam. Research-wise, Burstein and Chodorow (1999) found that the features considered by E-Rater are generalizable from native speaker writing to non-native speaker writing and that the system was not confounded by non-standard English structures. Leacock and Chodorow (2003) also claim that recent advances in automatic detection of grammatical errors are quite promising for learner scoring and diagnosis. In line with this idea, Lonsdale and Strong-Krause (2003), having explored the use of NLP for scoring novice-mid to intermediate-high ESL essays, claim that "with a robust enough parser, reasonable results can be obtained, even for highly ungrammatical text" (p. 66). Undoubtedly, much improvement is needed to construct automated scoring systems that would capture the distinctiveness of learner language, but this can be achieved by integrating a combination of scoring techniques, which will allow for building diagnostic models of learner writing. One approach to this is developing evaluation systems which target a set of well-defined constructs and compare the result of the input text with a corpus of similar previously analyzed texts. The output of the system can range from a simple comparison of the input text with the corpus to an elaborate explanation of what errors have occurred in the text and what steps could be taken to correct those.

CONCLUSION

Based on past work on automated scoring systems, it appears such systems that provide individualized feedback in a variety of ways to ESL writers is a goal that may be within reach. To date, very little work has been done in this area despite the technical capabilities currently available (Chapelle, 2006). In this paper we have discussed several successful automated scoring systems that have been developed recently; their use is rapidly growing, which can and should positively affect developments in computer-assisted language testing by prompting research on diagnosis, which in turn may help to develop our understanding of the variable underlying the development of writing proficiency.

The empirical research aimed at developing scoring systems through the use of NLP and statistical methods should provide much concrete evidence about writing development as

it is reflected in many aspects of learners' texts. Therefore, the insights gained from learner corpora used in automated systems for training purposes are relevant for this research agenda. In the long run, such an understanding could contribute to the formulation of a more specific writing proficiency framework than the ones that have been developed based on intuition and teaching experience.

This research also promises to provide data and experience that can inform theory and practice in diagnostic language assessment. As Xi (this volume) and Carr (this volume) show, automatic response scoring affects central issues in test design and validation. According to Jang (this volume), research is needed to assess the effectiveness of automated feedback. In short, "the potential of automated essay evaluation […] is an empirical question, and virtually no peer-reviewed research has yet been published that examines students' use of these programs or the outcomes." (Hyland & Hyland, 2006, p. 109).

Diagnostic writing tests need to develop from computer-based selected responses assessing recognition to automated systems-based assessment of written language production. We have attempted to justify our argument by pointing out the advantages of automated analysis of constructed responses and of automated feedback for developing learners' writing proficiency. However, we acknowledge that this venture is not an easy one. Designing an automated diagnostic writing test that satisfies all the necessary constraints will require a lot of incremental work. Because diagnostic tests "should be thorough and in-depth, involving an extensive examination of variables" (Alderson, 2005, p. 258), they should be creative in the use of NLP and statistical methods; therefore, close collaboration among specialists in computer science, computational linguistics, language assessment, CALL, and other related areas is needed to achieve desired results.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Alderson, J. C., Clapham, C. M., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

American Council for the Teaching of Foreign Languages (1983). *ACTFL proficiency guidelines* (revised 1985). Hastings-on-Hudson, NY: ACTFL Materials Center.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Baum, E. B. (2004). *What is thought?* Cambridge, MA: MIT Press.

Bejar, I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, *21*(2), 175-189.

Bennett, R., & Ward, W. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed responses, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Burstein, J. (2003). The E-rater scoring Engine: Automated Essay Scoring with Natural Language Processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL99 workshop on computer-mediated language assessment and evaluation of natural language processing*. College Park, MD. Retrieved August 3, 2007 from http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf.

Carroll, S. (2001). *Input and evidence: The raw material of second language acquisition.* Amsterdam: Benjamins.

Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, *15*, 357–366.

Chapelle, C. (2006). Test review. *Language Testing*, *23*, 544-550.

Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, and assessment*. Cambridge: Cambridge University Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: University of Cambridge Local Examination Syndicate and Cambridge University Press.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, *41*(6), 391-407.

Dikli, S. (2006). An overview of automates scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 4.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, R. (1994). A theory of instructed second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 79-114). San Diego, CA: Academic Press.

Heift, T. (2003). Multiple learner errors and meaningful feedback: A challenge for ICALL systems. *CALICO Journal*. *20*(3), 553-548.

Holland, V. M., Maisano, R., Alderks, C., & Martin, J. (1993). Parsers in tutors: What are they good for? *CALICO Journal*, *11*(1), 28-46.

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing, 7*(3), 255-286.

Hyland, K., & Hyland, F. (Eds.). (2006). *Feedback in second language writing: Contexts and issues.* New York: Cambridge University Press.

Keith, T. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-167). Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T., Laham, D., & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates.

Leacock, C., & Chodorow, M. (2003). Automated grammatical error detection. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195-207). Mahwah, NJ: Lawrence Erlbaum Associates.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 487-536). San Diego, CA: Academic Press.

Lonsdale, D., & Strong-Krause, D. (2003). Automated rating of ESL essays. *Proceedings of the HLT-NAACL 03 workshop on building educational applications using natural language processing, 2*, 61-67. Retrieved July 20, 2007 from http://ucrel.lancs.ac.uk/acl/W/W03/W03-0209.pdf.

Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *48*, 183-218.

Mislevy, R., Steinberg, l., Almond, R., & Lukas, J. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. Williamson, R. Mislevy, & I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, NJ:Lawrence Erlbaum Associates.

Mitchell, R., & Myles, F. (1998). *Second language learning theories*. London: Arnold Publishers.

Mitchell, T., Russel, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. In *Proceedings of the 6th CAA Conference*, Loughborough: Loughborough University. Retrieved July 31, 2007 from http://hdl.handle.net/2134/1884/.

Moussavi, S. A. (2002). *An encyclopedic dictionary of language testing* (3rd ed). Taiwan: Tung Hua Book Company.

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning, 50*, 617-673.

Nagata, N. (1998). The relative effectiveness of production and comprehension practice in second language acquisition. *Computer Assisted Language Learning*, *11*(2), 153-77.

Nagata, N. (1995). An effective application of natural language processing in second language instruction. *CALICO Journal*, *13*(1), 47-67.

Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal*, *77*(3), 330-9.

Ortega, L. (2007). Second language learning explained? SLA across nine contemporary theories. In B. VanPatten & J. Williams (Eds.) *Theories in second language acquisition: An introduction* (pp. 224-250). Mahwah, NJ: Erlbaum.

Page, E. (2003). Project Essay Grade. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Lawrence Associates.

Pawlikowska-Smith, G. (2000). *Canadian language benchmarks 2000: English as a second language for adults*. Ottawa: Citizenship and Immigration Canada.

Phillips, S. (2007). *Automated essay scoring: A literature review*. Society for the Advancement of Excellence in Education. Retrieved July 12, 2007 from http://www.saee.ca/pdfs/036.pdf.

Rosa, E., & Leow, R. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics*, *25*, 269-292.

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, *7*(26). Retrieved July 23, 2007 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/5e/46.pdf.

Rudner, L., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2), 3-21.

Scalise, K., & Bernard, G. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning and Assessment*, *4*(6), 4-43.

Schonell, F. J., & Schonell, F. E. (1960). *Diagnostic and attainment testing, including a manual of tests, their nature, use, recording, and interpretation.* Edinburgh: Oliver and Boyd.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Valenti, S., Nitko, A., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education* (Information Technology for Assessing Student Learning Special Issue), *2*, 319-329.

Van der Linden, E. (1993). Does feedback enhance computer-assisted language learning. *Computers & Education, 21* (1-2), 8161-65.

Williamson, D., Mislevy, R., & Bejar, I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wylie, E. &Ingram, D. (1995/99). *International second language proficiency ratings (ISLPR): General proficiency version for English*. Brisbane: Center for applied Linguistics and Languages, Griffin University.

Notes

[i] The frame is "somewhat detailed" considering that it was meant to inform item development for 14 languages covered by DIALANG. More details were added depending on the peculiarities of individual languages.

[ii] Alderson (2005) discusses grammatical categories when describing DIALANG's grammar test; however, we found this material very relevant in this context.

[iii] Automark also makes use of an information extraction approach, which is considered a shallow NLP technique as it typically does not require a full-scale analysis of texts.

[iv] Adjacent agreement is different from exact agreement in that it requires two or more raters to assign a score within one scale point of each other (Elliot, 2003).

Carr, N. (2008). Decisions about automated scoring: What they mean for our constructs. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 82-101). Ames, IA: Iowa State University.

Decisions about Automated Scoring:
What They Mean for Our Constructs

Nathan Carr
California State University, Fullerton

This paper discusses how decisions about the scoring criteria used in the automated scoring of constructed response items can affect the constructs that the test is intended to assess. It begins with a discussion of the benefits of automated scoring, followed by a brief overview of three general approaches to automated scoring—natural language processing (NLP), exact word matching, and keyword matching. The paper then focuses on the use of keyword matching in scoring comprehension items, and reasons why this approach is clearly superior to exact word matching, and can be preferable in some cases to the more powerful method of NLP. Using the classification scheme developed by Carr, Pan, and Xi (2002), the paper considers the ways in which decisions involving the implementation of automated scoring can affect the constructs of a test, dividing these effects into unintended alterations, purposeful/principled refinement, and mixed cases. It focuses on the effects of seven categories of decisions: exactness of responses, partial credit, "undesirable" responses, synonyms, paraphrases, misspellings, and penalizing for extraneous information. Examples from a reading comprehension test scored using keyword matching are provided in order to illustrate the potential results of various choices in each of the seven areas. This is accompanied by a discussion of ways in which to implement keyword matching approaches in the context of web-based testing (WBT), with an emphasis on low-budget approaches, including the author's ongoing development of a keyword matching automated scoring program which runs in Microsoft Excel (Microsoft Corporation, 2003).

INTRODUCTION

Although computer-based testing (CBT) has been an important area of focus in language testing since the mid-1980s (Chalhoub-Deville, 2001), it has not led to much improvement in the tests themselves; a continued reliance on multiple-choice items (Alderson, 2000; Chalhoub-Deville, 2001) has essentially led to little more than the mechanization (Canale, 1986) of paper-and-pencil test tasks. Laurier (2000) notes that for many years, the main exception to this has been computer-adaptive testing (CAT), which allows more efficient testing, and allows developers to specify the maximum amount of measurement error to be allowed (Hambleton, Swaminathan, & Rogers, 1991). The resources needed to develop a CAT appropriately—large item banks, and hundreds or even thousands of test takers—limit the spread of this type of testing, and generally keep its development out of the reach of any program that tests fewer than several hundred

students per year. Furthermore, CAT has been limited to selected response formats, generally multiple choice questions. This has primarily been because of the need for immediate scoring, although no doubt the traditionally strong association between item response theory—the measurement approach normally employed with CAT—has played a role here as well.

The long-hoped-for potential of CBT to innovate language testing may be on the verge of becoming realized, however, thanks to the introduction of Web-based testing (WBT). Its advantages include the potential to offer flexible delivery with customized formatting, centralized collection of responses, and elimination of the need for local installation of test software. These are not, however, the main benefits that it promises to deliver. It is the potential of WBT to make practical the automated scoring of constructed response tasks that is probably its greatest strength. Automated scoring in a WBT environment offers many benefits not available with other testing formats, including traditional CBT.

First and foremost among these is probably that automated scoring in WBT makes short answer responses practical, by eliminating the time, effort, and expense needed for human scoring. In any situation with more than a few test takers or a lengthy turnaround time between testing and score reporting, programs have heretofore been forced to rely upon selected response items; now, more authentic tasks that reduce the likelihood of successful random guessing can be adopted, once computer access bottlenecks are dealt with. Using automated scoring for constructed response tasks also improves the consistency of scoring, as a particular response will always receive the same rating or score. In the event that scoring criteria are changed, it is possible to rescore examinees' responses rapidly without any additional expense. This ease of rescoring also allows the consideration of alternative scoring decisions and comparison of their results. Finally, keys or other scoring criteria need to be in place before test administration if scoring is to be done in a timely fashion, and this advance specification seems to help lead to better test items. For example, constructed response items that seemed clear enough when written may actually prove difficult to score, and specifying the answer in advance can help to bring such items to the attention of test developers.

APPROACHES TO AUTOMATED SCORING

Approaches to automated scoring can be classified into three categories: natural language processing (NLP), exact match scoring, and keyword (also known as regular expression) matching. Although the last of these is the focus of this paper, the first two merit some discussion first.

NLP, which is also used for essay scoring, has been the focus of much of the research on automated scoring (Leacock & Chodorow, 2003; see, e.g., Higgins, Burstein, Marcu, & Gentile, 2004). NLP systems attempt to process a text to mimic understanding of the text and are probably the only way that extended production tasks can be computer rated when content is part of the scoring criteria. When linguistic accuracy is part of the construct, NLP systems can also be designed to assess this as well via analysis of the

linguistic features of responses (see, e.g., Li, 2000). NLP has certain drawbacks, however. In particular, it requires developing or licensing complex—i.e., expensive—software. Additionally, this software must then be "trained" using previously scored responses, typically numbering in the hundreds. Nevertheless, this appears to be the most promising approach for scoring essays, and perhaps even for assessing speaking as well (see Educational Testing Service, 2006; and Xi, 2007, this volume). It has also seen some use in scoring shorter responses, primarily in the Educational Testing Service's c-rater scoring engine, although these are generally long for limited production tasks, with responses averaging two to five sentences (Educational Testing Service, 2006).

Another approach to automated scoring is exact match scoring. As its name implies, the entire response must match the key exactly. Systems using this approach generally seem able to deal with extra spaces between words, and are not usually case sensitive, but this is the limit of their flexibility. Exact-match scoring is available through the quiz function of course management systems such as Blackboard (Blackboard, 2007). It is also used by CBT systems that only allow one-word responses in limited production responses (e.g., the Questionmark Perception system; Questionmark, 2007). This approach is only really practical for one-word answers, or perhaps short set phrases. Using it with sentence-length response is also possible, of course, but requires that the scoring key include every variant of every answer that will receive credit. This includes predicting in advance every potential misspelling of every word in the key, every acceptable synonym, and every acceptable arrangement of the correct words. For responses of more than one or two words, therefore, it is not really practical.

The third approach to automated scoring is keyword matching, sometimes referred to as regular expression matching, and this scoring method is the focus of this paper. Scoring engines employing keyword matching search examinee responses for the particular key word(s) specified by the test writers. This method works better than exact match scoring for limited production tasks with expected responses more than one or two words long, which makes it far more practical for reading or listening comprehension test questions. Naturally, although it is unlikely that test developers would want to do so, keyword matching can operate identically to exact word matching by specifying a single correct answer. More likely, however, a particular item might have only one correct answer. The software needed for scoring using keyword matching does not need to be nearly as complex as an NLP scoring engine, and does not require "training" on sample responses before it can be used—although reviewing the results from pilot testing for possible inclusion in the key may be prudent, as will be discussed below. Keyword matching requires careful specification and review of the scoring key, and this additional level of care has the potential to improve tests in generally unexpected ways.

POTENTIAL EFFECTS OF AUTOMATED SCORING ON CONSTRUCTS

Carr, Pan, and Xi (2002) divide the effects of automated scoring on test constructs into three types: unintended alterations, purposeful/principled refinement, and mixed cases.

The same framework will be used here to discuss the potential consequences of decisions arising from the implementation of automated scoring, particularly in the context of limited production constructed response tasks. In all three categories, the effects come about because the automated scoring system is the tool through which a test's criteria for correctness are operationalized.

Implementing automated scoring involves addressing a number of issues, some of which can result in the unintentional alteration of the constructs being assessed. Whenever any criteria other than those specified for the test are used, the constructs of the test are altered, in effect, whether test developers recognize this fact or not, and the test is no longer measuring exactly what it was intended to measure. As a result, then, any inferences drawn on the basis of such scores become questionable in proportion to the nature and degree of the alteration. Carr, Pan, and Xi (2002) identify two examples of this potential problem, involving how to handle spelling errors and paraphrased responses. Not designing the scoring engine to accommodate these two issues will result in scoring criteria being imposed other than those desired by test developers. How to handle both spelling errors and paraphrasing will be addressed in this paper as well, as there is no simple solution to either of them.

In contrast, in some cases the use of automated scoring can lead to purposeful or principled refinement of a test's constructs. Simply having thought everything through in advance in terms of what is an acceptable response to each item tends to strengthen the construct validity of score-based inferences. That is, all things being equal, test developers are far less likely to include items that are not clear, which should mean they are more likely to correspond directly to the item specifications, which are the operationalization of the construct definitions contained in the test blueprint, or test specifications for the overall test. For the same tasks, automated scoring is more likely than human scoring to lead to such refinement because of the greater degree of clarity required in specifying correct answers. The sort of imprecise specification that works well enough for humans embodied in "You know what I mean" and the terms "reasonable" or "like that" can often be clear enough for human scorers, but such imprecision does not yield favorable results with computers.

Carr, Pan, and Xi (2002) illustrate how such construct refinement can occur with the example of four low-level points in an incomplete outline task. The test development team had originally felt that the order in which the four responses were provided should not matter. The problem was that it proved impossible to score the four items separately without enforcing order and without combining them into a single item, unless test takers were to be allowed to give the same response in all four blanks. The purpose of the task was to test for sensitivity to rhetorical organization and the information structure of the passage; thus, upon reflection, it became clear that the order should, in fact, matter—even if that order is arbitrary, it is the order used in the passage.

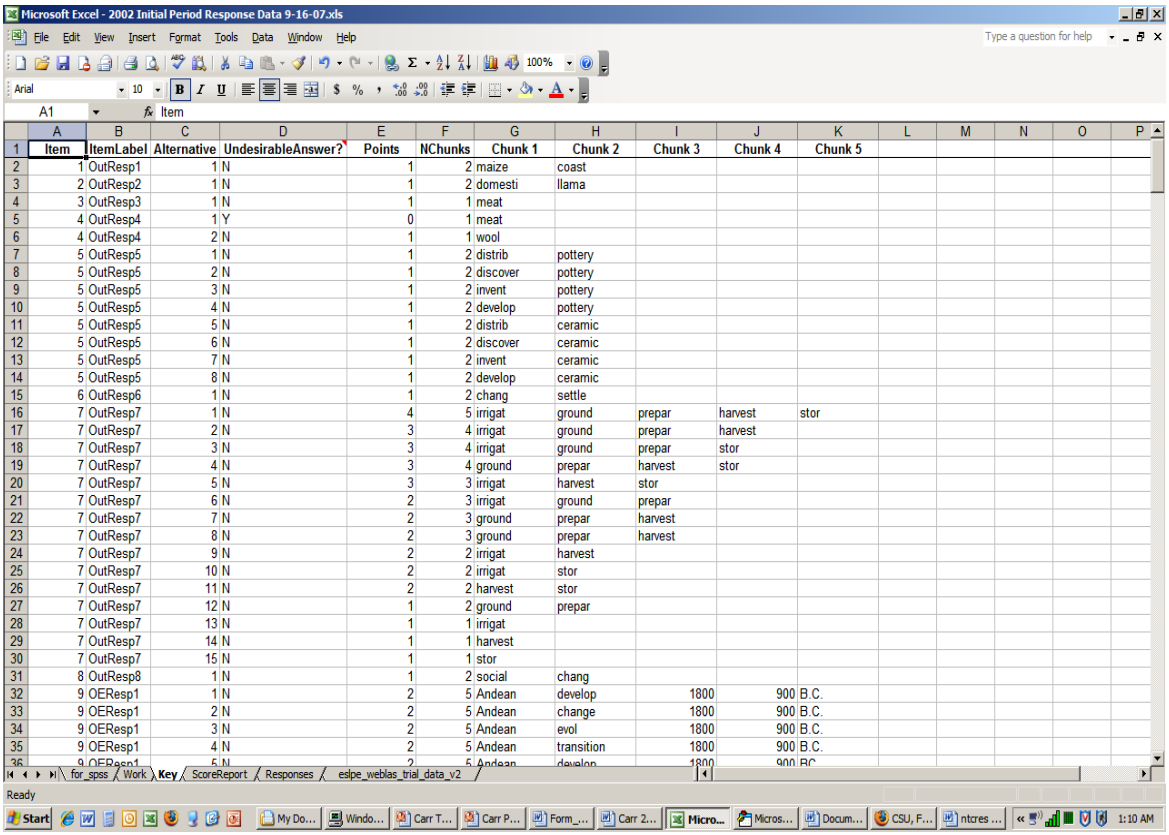
The mixed cases are best illustrated by the question of deciding how (and even whether) to penalize for extraneous information in responses. When a response includes extra,

unnecessary information, deciding to award it full credit, partial credit, or no credit means making a decision about what it is that is being assessed. The consequences of each possible decision must be weighed by test developers, as a given choice might lead to unintentional alteration of constructs, principled construct refinement, or a combination of the two. Given that there is no single right answer to the question of how to handle such responses, the issue of handling extraneous information must be addressed in each testing context, and is therefore discussed in greater detail below.

METHODOLOGY

This paper provides illustrations of the issues discussed using the same dataset that was used in Carr, Pan, and Xi (2002). That study, the findings of which are described in greater detail below, discussed ways in which automated scoring can alter the constructs assessed by a test, resulting in their principled refinement in some cases, and their unintentional alteration in others. The dataset includes 251 responses to an academic reading comprehension test developed at the University of California, Los Angeles for use as part of the university's ESL Placement Examination (ESLPE). The ESLPE is used to place incoming non-native speakers of English into the appropriate level of academic English instruction, and at that time included reading, writing, and listening sections. The portion of the test discussed here consisted of 11 incomplete outline items and 10 short-answer items. The responses were obtained in 2002 during pilot testing of these sections, which were not included in determining students' scores. This was done as part of the development of the Web-Based Language Assessment System (WebLAS), which was first used for operational testing in March 2006. The PoorMan Scoring System, developed by the author, was used here to rescore all responses and to analyze the effects of different decisions about scoring. The goal of the PoorMan system is to provide a low- or no-cost system (i.e., a poor man's scoring engine) for automated scoring that can process response data contained in a spreadsheet or other delimited file, initially for research purposes, and perhaps eventually for automated testing. Data collection is, of course, a separate issue, and test delivery systems can range from the simple, such as using HTML web pages with basic forms and simple scripting (see, e.g., Birnbaum, 2001), to the more complicated, such as using Flash for greater test security (see Carr, 2006, for a discussion).

PoorMan is essentially a large Microsoft Excel (Microsoft Corporation, 2003) macro. For those less familiar with what such programs can do, it may be better to view it as a Visual BASIC program that uses the Excel interface for data input and output. It works by reading in the scoring key from one worksheet (see Figure 1 for an example of a portion of a key), scoring the responses contained in a second worksheet, and then entering item-level scores in a third worksheet. At present, the key must be entered manually, including all acceptable synonyms. The next development step for the system will be the construction of a key generator, which will ask test writers for a model answer, prompt them to identify the key terms (usually one word each), and then ask them to accept and reject synonyms. Once entered, the key can be altered manually in Excel.



Microsoft Excel - 2002 Initial Period Response Data 9-16-07.xls

Item	ItemLabel	Alternative	UndesirableAnswer?	Points	NChunks	Chunk 1	Chunk 2	Chunk 3	Chunk 4	Chunk 5
1	OutResp1	1	N	1	2	maize	coast			
2	OutResp2	1	N	1	2	domesti	llama			
3	OutResp3	1	N	1	1	meat				
4	OutResp4	1	Y	0	1	meat				
4	OutResp4	2	N	1	1	wool				
5	OutResp5	1	N	1	2	distrib	pottery			
5	OutResp5	2	N	1	2	discover	pottery			
5	OutResp5	3	N	1	2	invent	pottery			
5	OutResp5	4	N	1	2	develop	pottery			
5	OutResp5	5	N	1	2	distrib	ceramic			
5	OutResp5	6	N	1	2	discover	ceramic			
5	OutResp5	7	N	1	2	invent	ceramic			
5	OutResp5	8	N	1	2	develop	ceramic			
6	OutResp6	1	N	1	2	chang	settle			
7	OutResp7	1	N	4	5	irrigat	ground	prepar	harvest	stor
7	OutResp7	2	N	3	4	irrigat	ground	prepar	harvest	
7	OutResp7	3	N	3	4	irrigat	ground	prepar	stor	
7	OutResp7	4	N	3	4	ground	prepar	harvest	stor	
7	OutResp7	5	N	3	3	irrigat	harvest	stor		
7	OutResp7	6	N	2	3	irrigat	ground	prepar		
7	OutResp7	7	N	2	3	ground	prepar	harvest		
7	OutResp7	8	N	2	3	ground	prepar	harvest		
7	OutResp7	9	N	2	2	irrigat	harvest			
7	OutResp7	10	N	2	2	irrigat	stor			
7	OutResp7	11	N	2	2	harvest	stor			
7	OutResp7	12	N	1	2	ground	prepar			
7	OutResp7	13	N	1	1	irrigat				
7	OutResp7	14	N	1	1	harvest				
7	OutResp7	15	N	1	1	stor				
8	OutResp8	1	N	1	2	social	chang			
9	OEResp1	1	N	2	5	Andean	develop	1800	900	B.C.
9	OEResp1	2	N	2	5	Andean	change	1800	900	B.C.
9	OEResp1	3	N	2	5	Andean	evol	1800	900	B.C.
9	OEResp1	4	N	2	5	Andean	transition	1800	900	B.C.

Figure 1. A portion of a scoring key.[i]

When performing the scoring, PoorMan begins by checking the key for format errors. The system then scores one item at a time (see Figure 2). Each test taker's response is searched for alternatives, that is, possible answers that are contained within the key. "Undesirable" alternatives (discussed below) are searched for first, and these are typically awarded zero points. If no undesirable alternatives are found in a response, it is then searched for alternatives worth full credit, then for those worth partial credit. If none of the alternatives in the key are found in a given response, it receives zero points. Once an alternative has been found within a given test taker's response, or if none of the alternatives are found in the current response, the system moves on to the next person. Once all of the test takers' responses to an item have been scored, the system moves on to the next item.

The alternatives from the key are searched for by looking for the "chunks"—that is, regular expressions, or key words—that comprise them. At present, PoorMan does not require these chunks to be in the order in which they appear in the key, except when one is a multiword expression; for example, "post" and "office" would be considered two key words, but "post office" would be treated as one "word" with a (mandatory) space in the middle.

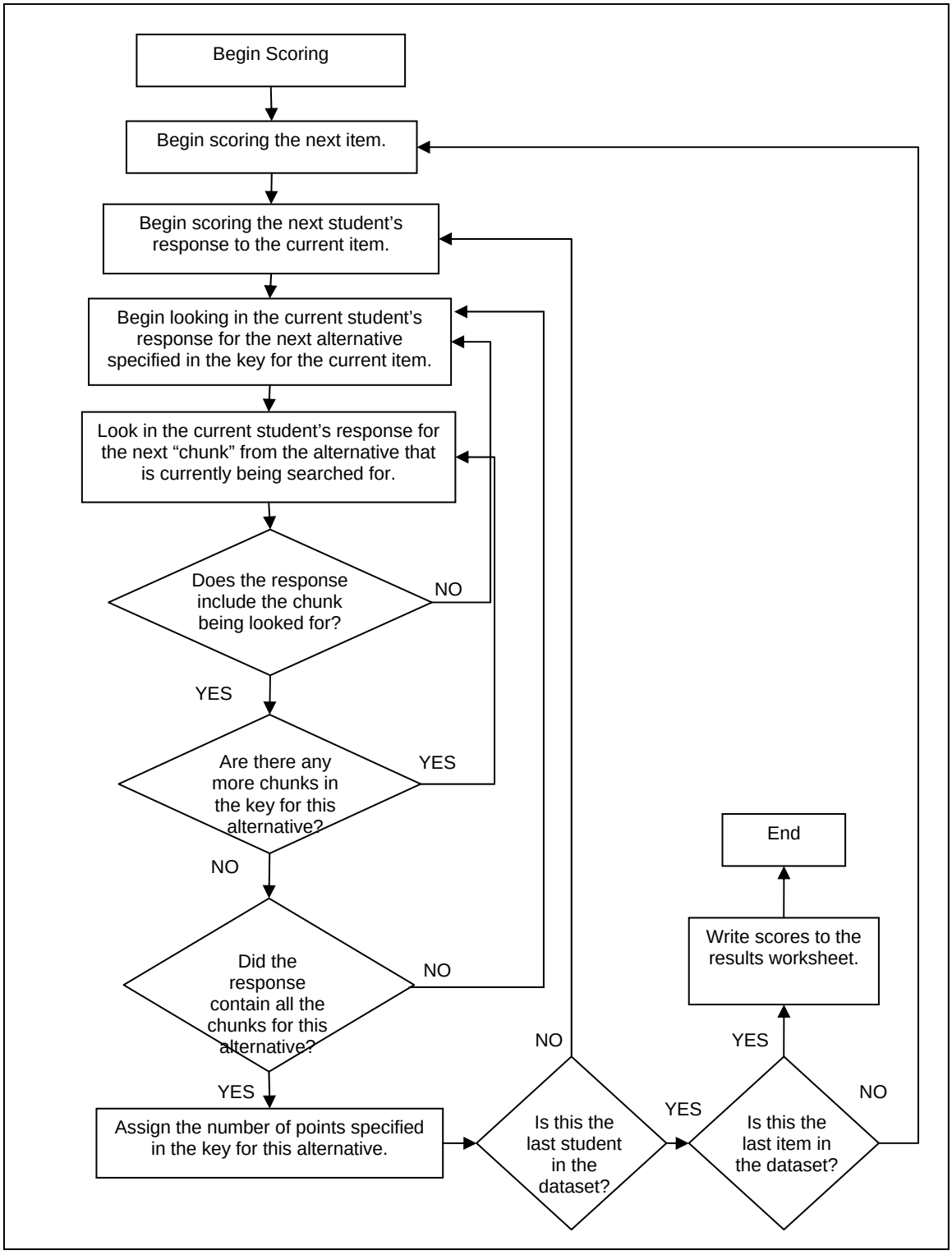


Figure 2. Flow chart illustrating the decisions made during the PoorMan scoring process.

The PoorMan engine yields results that are highly similar to those produced by the UCLA WebLAS scoring engine and referred to by Carr, Pan, and Xi (2002), although inspection of discrepancies indicates that at least some are attributable to differences between the keys, particularly in terms misspellings not yet added to the key being used with the PoorMan system. Additionally, there are cases in which it is unclear why the original WebLAS scoring engine counted certain responses wrong. For example, in an item where "maize along the coast" was the model answer and both *maize* and *coast* were required for credit, "maize,manioc,sweet potatoes,beans,peanuts, and other crops along the coast" was originally counted as incorrect, although the WebLAS engine *should* have treated it as correct, as PoorMan did. For another example, on an item with the model answer "domestication of the llama," where both *domestication* and *llama* were required for credit, "llama domesticated" was somehow considered incorrect by the original version of the WebLAS engine, but was accepted by the PoorMan scoring system. It appears, therefore, that the PoorMan scoring engine may represent an improvement over the early version of the WebLAS scoring engine used by Carr, Pan, and Xi.

MAKING DECISIONS ABOUT SCORING CRITERIA AND CONSTRUCTS

Test developers need to consider several issues, and make decisions in advance, if they are to maximize the construct-related benefits of automated scoring while avoiding unintended alterations. In particular, decisions need to be made in each of the following areas: exactness of responses, partial credit, "undesirable" responses, synonyms, paraphrases, spelling, and penalizing for extraneous information.

Exactness of Responses

Decisions about exactness of responses overlap with those involving partial credit, synonyms, paraphrases, and spelling, which are discussed below. Such decisions are affected by how much trouble we are willing and able to go to in creating the key, which is affected by the resources available. It most clearly emerges as a potential issue separate from the others discussed here in the context of key terms consisting of multiple words.

For example, in an item with the key[ii] *(have no)/(has no)/(had no) + container*, which requires two elements to receive credit (*have no container*, *has no container*, or *had no container*), developers might consider changing to a key with three required elements: *(have/has/had) + no + container*. Another example might be choosing between *sweet potatoes* and *sweet + potatoes.* The main advantage for the change in this case would be to accommodate misspellings of "sweet" (e.g., "sweets") or the insertion of additional spaces between the two words, as the scoring engine would be looking for the two words separately, rather than trying to find the two words together with a single space between them. In the case of the data used here, neither of these changes would change any examinees' scores; in another case, however, it makes a tremendous difference. For an item asking about one of the benefits of domesticating the llama in ancient Andean civilizations, the model answer—a direct quote from the passage—is "transport power." Using *transport power* or *transport + power*, the model answer, would have yielded an

item facility (IF) of .494 in this dataset; however, the test development team decided that recognizing that the llama provided benefits involving transportation was enough to indicate comprehension, and therefore specified *transport* as the key. This resulted in an IF of .777 instead.

Obviously, there is no one correct policy across all testing contexts for deciding what constitutes a sufficiently "exact" answer and what does not, and it would probably be difficult at best to articulate such a policy even for just one test. It is therefore essential for test writers and other members of test development projects to keep firmly in mind what they are attempting to assess with a particular item; anything that requires knowledge or abilities beyond that construct introduces construct-irrelevant variance (Messick, 1989). At the same time, they must avoid watering down their criteria for correctness; in the case of comprehension items, it must be clear to them that a particular answer *would* demonstrate comprehension. This process of consideration and evaluation clearly offers the potential, albeit not the promise, of enhanced construct validity.

Partial Credit

In general, including partial credit should be helpful in terms of improving overall test usefulness. Abeywickrama (2005) and Henning et al. (1993) report that partial credit scoring increases reliability, which is logical, as it allows each item to provide more information than would be the case with dichotomous scoring. Furthermore, Abeywickrama also reports that in confirmatory factor analysis, a gap-fill task provided better model fit than dichotomous scoring; this is probably best interpreted as indicating an improvement in construct validity. All this being said, however, the question remains of *how* to implement partial credit scoring. The issues here are by no means unique to automated scoring, but because it requires prior consideration of the scoring key, partial scoring tends to force a more principled, *a priori* consideration of how partial credit is to be awarded. This stands in marked contrast to the judgment-based, often *ad hoc* decisionmaking usually involved in human scoring.

When making decisions about partial credit, test developers must consider several things, starting with what the purpose or construct is of the test section and/or item type. They must determine exactly what information is being sought in that item, as well as how much of the model answer test takers need to provide to receive full or partial credit—in other words, to demonstrate full or partial mastery of the portion of the construct assessed by the item. In a comprehension test, for example, the point is to demonstrate comprehension; therefore, if the examinee has to demonstrate more than comprehension because of the way the scoring key is specified, the construct is altered. These types of decisions should lead to construct refinement on an item-by-item basis.

An illustration of how to award partial credit for partially correct responses can be seen in the case of an item asking about the main idea of the reading passage used in this study. The model answer, worth two points, was, "Between 1800 and 900 B.C., the way of life in Andean South America changed drastically." Test takers would receive one point for

answering *Andean + develop/change/evol/transition*: the word "Andean" along with any word containing "develop," "change," "evol," or "transition." Thus, any word starting with "develop," such as "development," or even "developing," or any word beginning with "change" (but not "changing," which has no *e*), or starting with "evol" or "transition," would be acceptable as long as it was accompanied by "Andean." Test takers would also receive one point for *1800 + 900 + BC/B.C.*, that is, for including both years along with either "BC" or "B.C." Providing both halves of the key would then yield a total of two points. Examples of student responses to this item, along with the points awarded for them, are provided in Table 1.

Table 1. Examples of Examinee Responses to a Question Worth up to Two Points

Points awarded	Examinee response
2	Explain how between 1800 and 900 BC, life change drastically in Andean South America
2	The development of life of Andean South America during 1800-900 BC.
2	Andean evolutionary developments between 1800 and 900 B.C.
1	Andean evolutionary
1	The life in Andean South America between 1800 and 900 B.C.
1	the rise of the first Andean States.
1	The changes brought from maize. It changed a civilization way of life.
0	The Early Andean Culture
0	theocrological chages in the ititial period.[a]
0	Ancient civilization in the Andean region of S. America.[b]
0	civilizations.
0	Tell the chages in the agriculture economy ad therefore the moving inland.

Note. The key Andean + develop/change/evol/transition is worth one point, and 1800 + 900 + BC/B.C. is worth one point, for a total of two points maximum possible credit.
[a]As this indicates, "chage" needs to be added as an acceptable misspelling of "change." Also, "Initial Period" might be considered for inclusion as an acceptable alternative to "Andean." [b]"Ancient" should probably be considered as an alternative to "between 1800 and 900 B.C."

A second example involves an incomplete outline item, and illustrates the effect that different decisions regarding what constitutes an acceptable—or partially acceptable—response can have. For this item, the model answer is "Changes in the distribution of settlements," which is a direct quote from the reading passage, and is worth one point. If *chang + settle* is given full credit, this yields an IF of .494, with 124 out of 251 examinees answering correctly. If credit is only awarded for responses more closely resembling the model answer (*chang + distribut + settle*), however, IF is only slightly reduced to .484, and 119 examinees are considered to have answered correctly. On the other hand, awarding .5 points for *chang + settle* yields an IF of .484. These differences seem trivial at first glance, but this may be in part because the scores in question are not ours. Furthermore, .5 points constitutes 4.5% of the total points for the incomplete outline task, which may help to put the matter into clearer perspective. Most examinees would probably consider decisions worth 4.5%, let alone 9%, to be fairly important.

"Undesirable" response

Another area in which decisions must be made, and where automated scoring can prompt test developers to refine their constructs and thereby enhance the construct validity of their tests, is in deciding what constitutes an "undesirable" response. Undesirable responses are those that are the opposite of the correct response, or otherwise deviate from the model answer sufficiently to indicate to a human rater that the examinee did not, in fact, provide a correct answer. They differ from ordinary wrong answers in that they contain enough keywords to receive full or partial credit from the automated scoring system.

Often, such answers are caught by the scoring engine without the need for special treatment, but not always. An example is a question asking where most Andean pyramids had been built prior to the period discussed in the reading passage. The exact wording in the passage was "close to the shore," and since "on the coast" was judged an acceptable paraphrase as well, and the prepositions were deemed not essential to demonstrating comprehension, the key was specified as *shore/coast*. One undesirable answer added to the key (Cerro Sechin, an inland location mentioned in the passage) had no effect on anyone's scores, as none of the test takers in the sample provided it as an answer. On the other hand, adding *inland* as an "undesirable" led to five responses being counted incorrect that had previously been accepted by the scoring system. Including *far/away + shore/coast* identified one additional response that had been counted correct. It is therefore important that test writers give consideration to incorrect responses containing the appropriate keywords, or other terms that would indicate a lack of comprehension. Doing so constitutes an incremental refinement of the construct(s) being assessed; failing to do so, on the other hand, introduces construct-irrelevant variance—arguably random[iii] error variance, which also reduces the reliability of the test—by allowing some examinees to receive credit for an item who should have received none.

Synonyms

The question of synonyms must also be addressed when implementing automated scoring. Reasonable synonyms must be accepted if unmotivated construct alteration is to be avoided. One approach to dealing with this issue is to identify synonyms during pilot testing that test writers did not think of when creating the key. For example, the original model answer to an item asking how the word *remains* was used in the passage was "things left behind," or "something left behind." The key was therefore *thing/something + left*. Following pilot testing, however, 10 additional examinee responses were identified as being acceptable, leading to full credit for *remnants* or for *(thing/something/what/whatever) + (left/leav/(can + found/find))*. While somewhat complicated, it poses no problems for the scoring engine once added, and allows the system to accommodate a wider range of responses that a human rater *should* find acceptable.

Merely relying on a review of pilot testing data to find acceptable synonyms is probably not satisfactory. It is therefore worth considering another additional approach, including a thesaurus in the key generation module, or consulting one (either electronic or hard-copy) while creating the key manually. This should leave fewer desirable answers slipping through the cracks, but it lengthens the key creation process. Furthermore, a thesaurus often contains *numerous* poor synonyms, each of which must be rejected individually.

One problem with both approaches is that they introduce the risk of almost ridiculously long keys. This is illustrated in the key to a question about changes in eating habits. In the key, *(increase/more) + (maize/manioc/sweet potatoes/beans/peanuts/crop)* was worth one point, and *(decrease/few/less) + (fish/shellfish/littoral/forage/seafood)* was also worth one point, for a possible total of two points. For the first half, there are 12 possible combinations of alternatives (6 x 2 = 12) which would be worth one point, and there are 15 for the second half. For the entire item, therefore, there are 180 possible combinations of alternatives that are worth two points, and 27 that are worth one point. While they only slow down the scoring by a matter of seconds, such lengthy keys can pose a problem in terms of their creation. They are less troublesome, of course, if they are created in conjunction with a system that partially automates the process; on the other hand, when keys are created manually, and every acceptable response must be entered separately, the process can take quite a long time. In response to this, test developers may attempt to word items more carefully in order to reduce the number of acceptable synonyms, which would arguably be a case of construct refinement, or at least an enhancement of construct validity. On the other hand, more stringent scoring might result in other cases, which would be a matter of unintended construct alteration. The issue of how to handle synonyms, then, has the potential to go either way, and therefore requires caution on the part of developers and writers.

Paraphrases

As with synonyms, failing to accommodate reasonable paraphrases will alter constructs for the worse; worse still, it may penalize the strongest test takers, as these are probably the ones best able to express concepts in their own words. Even more so than with

synonyms, however, there is the substantial problem that not every reasonable answer can be anticipated. Review of responses following pilot testing, or even first operational testing, is therefore essential. This can be the difference between zero and full credit; for example, in an item with "Carvings and sculpture at Cerro Sechin seem to be scenes of soldiers being killed, trophy heads, and other similar designs" as the model answer, adding *scene* as an alternative to *carv/sculpt* (for "carvings" and "sculptures") in the key for one item resulted in two examinees receiving two points, rather than none. Two out of 251 is fairly inconsequential, of course—unless you are one of those two test takers.

A second example involved a question asking why it was probable that ancient Andeans had developed ceramics by themselves. The model answer, "the need/importance of containers for agriculture," was worth one point. Another response identified following pilot testing was "the transition to agriculture." This answer seems to indicate comprehension, so it should also receive credit; including it changes its IF from .359 to .390, which means that eight more test takers received credit for the item.

Spelling

If spelling is not intended to be part of the construct, then failing to account for reasonably comprehensible spelling errors—particularly in the case of reading or listening comprehension—will result in unintended construct alteration. It is therefore regrettable that there do not appear to be any "magic bullets" at present for addressing this issue, particularly in a low-cost fashion that does not involve extensive professional programming. I have been able to identify six imperfect approaches, however, several of which in combination would probably do an adequate job of addressing the problem. The six options are to (1) tell test takers to be careful and proofread their responses; (2) spell-check all responses; (3) use shorter keywords; (4) use intuition and personal experience to predict common misspellings when creating the key; (5) use a dictionary of common misspellings or a typo generator when creating the key; and (6) review incorrect and partial-credit responses after pilot testing and add acceptable misspellings to the key.

Tell test takers to be careful. The simplest and cheapest option to implement in dealing with spelling errors is to tell test takers in the instructions to check the spelling in their answers, using the passage as a guide. This option is probably the least effective, too, however: Presumably, examinees are being as careful as they can already. However, it never hurts to encourage caution, so this method should be used, but not relied upon.

Spell-check all responses. Another approach is to spell-check all responses. Doing so manually, however, and requiring a human rater to accept or reject each word not found in the spell-checker's dictionary would partially defeat the purpose of automated scoring. It would appreciably add to the length of time needed to score the responses, and would require the establishment of standardized criteria for accepting and rejecting misspelled responses. On the other hand, using an entirely automated spell-checker would require a fairly sophisticated level of programming—since an outside program would have to be controlled by the scoring engine—and would therefore probably be rather expensive. Furthermore, aside from the technical difficulties involved in mating a spell-checker to

the scoring engine, there is the additional issue that most spell-checkers focus on phonologically-based errors, not typographical ones. It is also possible that spell-checkers created for use with native English speaker's writing might be problematic for use with the writing of non-native speakers.

Another problem with spell-checking all responses is that many words will be checked that are not important; for example, in the response data used in this paper, an experiment with manual spell-checking on one item produced a number of "hits" for the word "paralled." This presented two problems. First, "paralleled" was not a keyword being searched for in the examinee responses, so its correct spelling did not really matter. Second, "paralleled" should not even have been part of the answer to begin with. Despite this, it still had to be checked. Therefore, unless spell-checking can be automated, and can operate with a satisfactory degree of reliability, it will probably remain relatively impractical compared to other approaches.

Use shorter keywords. Using shorter keywords presents fewer opportunities for spelling or suffixation errors to matter, since fewer letters are being checked. Some examples would be "potatoes" vs. "potato" vs. "potat," or "settlements" vs. "settle" vs. "settl." This is probably one of the most technically facile approaches that can be used, and it is likely to take care of a fairly large number of spelling errors.

***Use intuition and personal experience to predict misspellings*.** To a certain extent, item writers and other members of the test development team can attempt to use their own intuition and personal experience to predict certain common misspellings. For example, if the key includes the word "distribution," writers might expect based on their own typing misadventures that "distirbution," "distributino," and "distributuon" might occur in the responses. Not everything can be predicted in this manner, however, and furthermore, this method has the potential to occupy a disproportionate amount of developers' time. Furthermore, if the misspellings are being manually entered into the key, this will require even greater amounts of time.

Use a misspelling dictionary or typo generator when writing the key. Manually looking up potential misspellings, whether in an electronic or hard-copy dictionary, is another option that would probably be fairly effective, but the amount of time this process would require would likely be rather onerous, particularly if the key were being entered manually, one alternative at a time. This approach is therefore probably most practical if the dictionary or typo generator is integrated with the scoring engine's key generation module. Judging from the options that they commonly list for generating typos (see, e.g., TheDowser Software, 2007; Wall, n.d.), such generators typically look for duplicate characters, swapped letters or characters, missing letters, extra letters, keyboard proximity errors, missing spaces, and phonetic errors. They can also apply custom rules in some cases.

Using a typo generator, however, requires the key writer to wade through a morass of plausible, implausible, and semi-plausible options, or accept *all* of them and thereby risk

overfilling the key. This is not an idle worry, even though Excel supports up to 65,536 rows in a given spreadsheet. As an illustration, consider an item discussed previously which had as its model answer "changes in the distribution of settlements," which contains three keywords ("changes," "distribution," and "settlements"). Using the Seobook typo generator (Wall, n.d.), these three keywords have 134 spelling variations, 260 variations, and 223 variations, respectively, for a total of 7,769,320 permutations. Using the searchspell (sic) typo search (Searchspell, 2000) yields a more manageable 15, 55, and 61 variations, with a total of "only" 50,325 permutations. As PoorMan is currently constructed, that would require a separate row for each permutation. If *chang + settle* is awarded partial credit, then a further 915 (15 x 61 = 915) to 29,882 (134 x 223 = 29,882) rows, depending on the generator used, would be required.

Even reducing the terms checked by the typo generator by only inputting the key terms, not the actual words from the model answer, only helps some: For *chang + distribut + settle*, searchspell produces 6, 19, and 27 alternatives, for 3,078 full-credit permutations and 162 partial-credit permutations. Seobook returns 89, 200, and 127 alternatives, totaling 2,260,600 full-credit permutations and 11,303 for partial credit. Even using a realistic key rather than the model answer, therefore, means that at most only about 20 items similar to this example could be accommodated per test. This further assumes that no synonyms or paraphrases would be allowed, which is hardly a reasonable assumption. Including all of the alternative spellings suggested by typo generators or misspelling dictionaries would therefore require revising the entire data structure used for the key.

Review responses from pilot testing for additional acceptable misspellings. While reviewing responses that were counted as incorrect or that only received partial credit will not identify all the spelling errors that might occur during operational testing, it is likely to identify many of the more common ones, assuming the pilot testing is conducted with an adequate number of examinees. This review could be combined with reviewing examinee responses for acceptable synonyms, thus saving some time. The process would probably be greatly facilitated if an additional module can be added to the scoring program that will separate the incorrect and partially correct answers and present them for review, preferably eliminating duplicates as it does. One example of how this can benefit test takers involves a two-point item which required examinees to include the dates 1800 and 900 B.C. Manual review of the examinees' responses appeared to indicate a problem with the scoring engine at first, until it was realized that three test takers had written "B.C" instead of "B.C." or "BC". The one point accounted for 7.7% of the total points possible on that section of the test.

Penalizing for extraneous information

Whether and how to penalize when test takers include extraneous information remains perhaps the most intractable issue discussed here, probably because of the subjective nature of identifying such responses. An example of one is a response to a question asking why it was likely that ancient Andeans had developed ceramics for themselves. The model answer was "the need/importance of containers for agriculture" or "the

transition to agriculture." One test taker's response, however, was that "Archaeology has no record of people who, having made the transition to agriculture with a concurrent need of containers, failed to learn how to make them." Unfortunately, this answer was copied and pasted directly from the passage. Carr, Pan, and Xi (2002) identify three approaches to dealing with this issue: assigning score penalties to particular pieces of extraneous information anticipated, imposing maximum length limits on responses, and penalizing responses over a certain length.

The first approach, assigning score penalties to particular pieces of extraneous information anticipated by test developers or identified during trialing, is fairly easy to implement by treating them as undesirable responses, as described above. The two largest problems here are that not everything will be predictable, and not everything will show up during trialing. It also does not address the inclusion of excessive, extraneous information *per se*. Nevertheless, as it would be at least somewhat effective, it should probably be part of the solution in most contexts.

Imposing maximum length limits on the text box used for responses, or not allowing the submission of responses exceeding the length limit, is another promising approach. In most cases, it should be able to prevent test takers from providing "kitchen sink" responses in which they copy and paste large chunks of text. On the other hand, it might inconvenience students who do a lot of paraphrasing. It requires test developers to strike a balance between establishing a reasonable limit that is not so small that it interferes with students' good faith attempts at answering in their own words, but at the same time is not so large that it is rendered ineffective. Another problematic issue for length limits stems from the fact that they involve altering the examinees' responses themselves, as opposed to the way in which they are scored. This arguably renders the task somewhat less authentic, as few real-life tasks would strictly prohibit answers over a certain length. Furthermore, as such "kitchen sink" responses might be interpreted as a sign of a *lack* of comprehension, allowing them and then penalizing examinees appropriately—whatever "appropriately" means—might be more desirable.

If length limits prove impracticable for some reason, or are judged to be an indicator of a lack of comprehension, then warning students not to simply regurgitate chunks of text—and then assigning score penalties if their responses exceed a certain length—may prove an effective alternative. The most appropriate approach would probably be to base the limit for each item on the length of model answer, but that leaves open the questions of how *much* longer is too long, and what fraction of a point should be deducted per word, character, or proportion of model response length over the limit. Naturally, decisions about length limits and penalties must be made without unintentionally changing constructs—in particular, verbosity should not be accidentally conflated with a lack of reading comprehension.

All things considered, a combination of the three approaches outlined here seems the most promising way forward at present. Penalizing specific responses will address certain common, specific problems on an *ad hoc* basis. Developers then need to choose between

imposing a maximum length limit on the text box used for responses; penalizing excessively long responses; or imposing a large length limit, and then penalizing for responses that are shorter than that, but still unreasonably long.

CONCLUSION

As mentioned previously, the web-based scoring system and the constructed-response test tasks that it made feasible entered operational testing use at UCLA in March of 2006. Other technical issues required greater attention, however, and as a result, a spell-checker was not added to the system, and no mechanism was ever instituted to penalize examinees for extraneous information. A length limit *was* imposed, though, for items that seemed more likely to elicit lengthy responses. Paraphrases and misspellings were handled by predicting likely-seeming alternatives, and by reviewing the responses obtained during pilot testing (Sunyoung Shin, personal communication, November 5, 2007).[iv]

This paper has attempted to delineate several areas that require advance consideration when implementing automated scoring. These categories of decisions involve the exactness of responses, partial credit, "undesirable" responses, synonyms, paraphrases, misspellings, and penalizing for extraneous information. Each of these seven interrelated areas needs to be considered separately; at the same time, however, some approaches to dealing with these issues can be applied to several of them, in some cases simultaneously (e.g., reviewing responses not receiving full credit for acceptable paraphrases, synonyms, and misspellings). The most important thing when making these decisions is that test developers remain focused at all times on the construct that they are attempting to assess. Dealing with problems on an *ad hoc* basis, without considering the ramifications of decisions, may lead to construct "drift," as unintended alterations creep into the way in which the constructs are operationalized. Test developers must avoid not seeing the forest for the trees—as any hiker can confirm, seeing too many trees and not enough forest is a good way to get lost in the woods.

Finally, research is still needed in several of these areas, including the question of how effectively some of the approaches to dealing with these issues will work, particularly in the cases of paraphrases, synonyms, and misspellings. Another area which should be investigated further is the comparative effects of different rules for penalizing responses that exceed length restrictions, and the extent to which verbose responses that indicate comprehension might be treated by the system as merely having been copied from the passage.

REFERENCES

Abeywickrama, P. S. (2005). Validating a scoring rubric for a rational deletion gap-fill test. Unpublished Ph.D. qualifying paper, University of California, Los Angeles.

Alderson, J. C. (2000). Technology in testing: The present and the future. *System, 28* (4), 593-603.

Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet.* Upper Saddle River, NJ: Prentice Hall.

Blackboard. (2007). *The Blackboard Academic Suite.* Retrieved October 20, 2007, from http://www.blackboard.com/products/Academic_Suite/index

Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (Ed.), *Technology and language testing* (pp. 29-44). Washington, DC: Teachers of English to Speakers of Other Languages.

Carr, N. T. (2006). Computer-based testing: Prospects for innovative assessment? In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching* (pp. 289-313). San Marcos, TX: CALICO (Computer Assisted Language Instruction Consortium).

Carr, N. T., Pan, M., & Xi, X. (2002, December). *Construct refinement and automated scoring in Web-based testing.* Symposium paper presented at the 24th Annual Language Testing Research Colloquium, Hong Kong.

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning & Technology, 5*(2), 95-98. Retrieved October 21, 2007, from http://llt.msu.edu/vol5num2/Deville/default.html

Educational Testing Service (2006, Fall). What's inside: New technologies improve automated scoring. *ETS Innovations, 1*(2), 3-5.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Henning, G., Anbar, M., Helm, C. E., & D'Arcy, S. J. (1993). Computer-assisted testing of reading comprehension: Comparison among multiple-choice and open-ended scoring methods. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium: Dedicated in memory of Michael Canale.* Alexandria, VA: Teachers of English to Speakers of Other Languages. Inc.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Annual Meeting of HLT/NAACL*, Boston, MA. Retrieved October 19, 2007, from http://ftp.ets.org/pub/res/erater_higgins_dis_coh.pdf

Laurier, M. (1999). The development of an adaptive test for placement in French. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 122-135). Cambridge, UK: Cambridge University Press.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short answer questions. *Computers and the Humanities, 37*(4), 389-405.

Li, Y. (2000, January). Assessing second language writing: The relationship between computerized analysis and rater evaluation. *ITL, 127-128*, 37-51.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Microsoft Corporation. (2003). *Microsoft Office Excel 2003* [Computer software]. Redmond, WA: Author.

Questionmark. (2007). *Questionmark–Windows Based Authoring–Question Types.* Retrieved October 20, 2007, from http://www.questionmark.com/uk/perception/authoring_windows_qm_qtypes.aspx

Searchspell Inc. (2000). *Searchspell typo generation tool.* Retrieved October 21, 2007, from http://www.searchspell.com/typo/

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: SAGE Publications.

TheDowser Software. (2007). *Misspelled keywords.* Retrieved October 21, 2007, from http://www.thedowser.com/misspelled-keywords

UCLA Center for World Languages (2007). *I Want to Register for the ESL Placement Exam.* Retrieved November 16, 2007 from http://www.international.ucla.edu/languages/esl/article.asp?parentid=29246

UCLA Department of Applied Linguistics and TESL & Center for Digital Humanities. (2003). *WebLAS (Web-based Language Assessment System).* Retrieved September 17, 2007, from http://www.weblas.ucla.edu/

Wall, A. (n.d.) *Typo generator.* Retrieved October 21, 2007, from http://tools.seobook.com/spelling/keywords-typos.cgi

Xi, X. (2007). What and how much evidence do we need? Critical considerations for using automated scoring systems. *Proceedings of the Fifth Annual Conference on Technology for Second Language Learning*. Available from http://www.public.iastate.edu/~apling/TSLL/5th_2007/proceedings2007/contents.html

[i] Microsoft product screen shot(s) reprinted with permission from Microsoft Corporation.
[ii] In this paper, entries from the scoring key are italicized, while model answers or words from answers are put in quotation marks. Wildcards are not marked by asterisks, but essentially any element of the scoring key functions as a wildcard because of the way in which the PoorMan scoring engine searches for those elements within an examinee's response.
[iii] One could argue that it is random error variance if examinees' mistakes were random, and they just happened to provide a response containing a keyword, rather than other words. On the other hand, this should be viewed as systematic error variance if something about the text or item predisposed certain test takers to provide that response—this would be a *p* x *i* interaction effect in generalizability theory (Shavelson & Webb, 1991) terms, and conceptually separate from random error variance, even in a *p* x *i* design, where the two could not be differentiated from each other.
[iv] In September 2007, the ESL Placement Examination became a composition-only exam (UCLA Center for World Languages, 2007). As a result, the web-based portions of the test have been shelved indefinitely.

Xi, X. (2008). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 102-114). Ames, IA: Iowa State University.

What and How Much Evidence Do We Need?
Critical Considerations in Validating an Automated Scoring System

Xiaoming Xi
Educational Testing Service

Building on Clauser, Kane and Swanson (2002), this paper illustrates how an argument-based approach can be applied to the validation of the TOEFL® iBT Speaking test which uses an automated scoring system called SpeechRater v.1.0. The paper outlines assumptions pertaining to the links between each stage in the score interpretation and decision making process. Finally, evidence needed to reject potential rebuttals against the inferences is described. By outlining the inferences underlying score interpretation, the paper shows the connections among various aspects of validity evidence and offers insights into practical issues arising in a validation process such as prioritization of different types of evidence.

INTRODUCTION

The emergence of automated scoring systems in the past two decades (see a review in Yang, Buckendahl, Juszkiewicz, & Bhola, 2002) has been accompanied by theoretical work that defines the nature and scope of validation and empirical research to validate these systems. Previous validation work has followed a piecemeal approach and addressed one or more of these three areas: (1) demonstrating the correspondence (in both agreement and reliability) between scores produced by automated scoring systems and by human scorers, (2) examining the relationship between automated scores and scores on external measures, and (3) understanding the scoring processes that automated scoring systems employ (Yang et al., 2002). These different areas of investigation could potentially contribute to an argument for using automated scoring in an assessment; however, a mechanism is needed to tie them together in a coherent manner. This mechanism should allow practitioners to determine the critical evidence needed in view of the targeted use of the automated scores, and to integrate and evaluate existing evidence to support an argument for using automated scoring in a particular learning or assessment context.

Fortunately, we have seen a few attempts in the last ten years to integrate automated scoring into the overall assessment process, or the overall validity argument for an assessment. The body of work described by Bennett and Bejar (1998) provides useful guidance for developing a valid computerized assessment for which automated scoring

has been planned from the outset. It also unveils the complexity of validation work related to automated scoring. Another body of work, initiated by Clauser, Kane and Swanson (2002), is most useful in guiding the development and synthesis of evidence to support the proposed interpretation and use of scores produced by an automated scoring system. Their approach integrates the various areas of validation reviewed in Yang et al. (2002) into a coherent argument and extends these areas to include decisions based on automated scores and consequences incurred from using automated scoring. It thus provides a working framework for weaving automated scoring into the validity argument for the whole assessment.

Building on Clauser et al. (2002), this paper illustrates the application of an argument-based approach to the validation of SpeechRater v1.0, an automated scoring system deployed for the TOEFL® Internet-based Test (TOEFL iBT) Speaking Practice test. By contextualizing the approach in a real-world application, it offers practical insights into how to prioritize the different types of evidence gathered to support validation research in light of the intended use of SpeechRater in an on-line practice environment.

CONCEPTUAL VALIDATION FRAMEWORKS FOR AUTOMATED SCORING

Bennett and Bejar (1998) noted that previous research has largely examined automated scoring in isolation from the other components of an assessment and contended that it should be seen as an integral part of the whole assessment process. While automated scoring is constrained by other aspects of the assessment process, automated scoring itself has influence on decisions with respect to other aspects of the assessment, such as construct definition, test and task design, test taker interface, and reporting methods. Bennett and Bejar proposed that the development of an automated scoring system should involve two key steps: 1) extracting and implementing relevant features, each of which evaluates an aspect of the performance; 2) combining them into a score that indicates the overall quality of performance. Further, these two steps could be manipulated to maximize construct representation and to improve the relationships between automated scores and human scores on the same test or on external criterion measures. Bennett and Bejar's conceptual approach is most useful in driving the development of a valid computerized assessment that involves automated scoring. By seeing automated scoring as a dynamic component in a computerized assessment system consisting of interrelated components, this framework emphasizes the importance of evaluating the scoring mechanism in the context of a validity argument for the assessment. It has thus broadened the scope of validity investigations regarding automated scoring. Although the relationship of automated scoring to the overall validity argument of the assessment is not emphasized, their paper provides a foundation for the subsequent work that shifts the focus to the complete validity argument.

Based on a critical analysis of empirical validation efforts on automated scoring systems, Yang et al. (2002) proposed a validation framework that they claim to be essentially an

elaboration of the one developed by Bennett and Bejar (1998). Using this broadened validation framework as the reference point, they noted two gaps in the existing literature. The first one was the dearth of literature that conceptualized potential threats posed by the use of automated scoring for construct relevance and representation. They also highlighted the point that the consequences of using automated scoring systems should be examined as part of a validity argument, which would include an investigation of the extent to which it affects the user's perceptions of the assessment and the way they interpret and use the scores. Although not explicitly discussed in their paper, the impact of automated scoring on teaching and learning, depending on the goals of a particular assessment, seems to be a natural expansion of the scope of consequences which is part of a validity argument (e.g. Kane, 2006).

Clauser et al. (2002) provided the most comprehensive and in-depth analysis of validity issues involved in automated scoring systems for performance-based tests, following a general argument-based approach to validating a whole assessment (Kane, 1992; 2001; 2002; 2006; Kane, Crooks & Cohen, 1999). With this approach, validation involves two stages: developing an interpretative argument and evaluating a validity argument. In the first stage, for each intended use of test scores, an interpretive argument is articulated through a logical analysis of the chain of inferences linking performance on a test to a score-based decision, and the assumptions upon which these inferences rest. The second stage involves an evaluation of the plausibility of the interpretive argument within a validity argument using theoretical rationales and empirical evidence.

This approach has not expanded the scope of validity investigations beyond that of Messick (1989), but its major strength lies in providing a transparent working framework to guide practitioners in three areas: prioritizing different lines of evidence, synthesizing them to evaluate the strength of a validity argument, and gauging the progress of validation efforts. This approach also allows for a systematic way to consider potential threats to the assumptions and inferences and allocate resources to collect evidence to discount or reduce the impact of such threats. In applying this framework to automated scoring, Clauser et al. (2002) discussed how decisions made in developing an automated scoring system may strengthen the overall validity argument or potentially weaken it, given the particular approach used to develop the system. Their discussion focused on the potential threats to the strength of each inference in the chain that may be introduced by automated scoring, pointing to the critical areas of research that are needed to discount or reduce the threats. Although Clauser et al. (2002) may not cover all the potential validity issues introduced by automated scoring, they provide a working model for integrating automated scoring into this network of inferences leading to the intended interpretation and use of test scores.

Their working model is used as a basis in this paper for examining issues that might impact the validity of the TOEFL iBT Speaking Practice test which uses SpeechRater v1.0. In addition, it identifies the most critical inferences to be supported given the purpose of the assessment and summarizes evidence that is needed to reduce the impact of the potential threats to each inference.

THE TOEFL iBT PRACTICE ON-LINE ASSESSMENT

SpeechRater v1.0 is intended to provide instant score feedback on the TOEFL iBT Speaking Practice test. This section provides a brief overview of the purpose of the TOEFL iBT Practice assessment and the tasks and scoring rubrics of the TOEFL iBT Speaking Practice test.

The TOEFL iBT Speaking Section is designed to measure the academic English abilities of non-native speakers who plan to study at English-medium institutions for higher education. The TOEFL Practice On-line (TPO) has been made available to help prospective TOEFL iBT examinees become familiar with and better prepared for the TOEFL iBT test. Using retired operational TOEFL iBT test forms, TPO is designed to mirror the content and design characteristics of the TOEFL iBT test to the extent possible. However, unlike the TOEFL iBT test, the TPO allows users to customize their practice and take the test in a timed or untimed mode. The timed mode attempts to replicate the operational testing experience by using the same on-line delivery system and timing restrictions of TOEFL iBT. In the untimed mode, users can progress at their own pace, starting or stopping the test whenever they like and revisiting items they have completed if desired. Another important distinction between the TPO and the TOEFL iBT test is that the former allows users to receive immediate feedback on their performance to help them assess their own comfort with the TOEFL iBT test administration. In early 2006 the users of TPO were able to instantly receive scores on reading and listening sections, both comprised of multiple-choice items that are computer scored, as well as the writing section, with automated writing scores provided by e-rater® (Attali & Burstein, 2005). The scores on speaking sections were produced by human raters within five business days. As a result of substantial interest in more immediate feedback from the speaking section of the TPO, a research agenda was launched to develop and deploy an automated system for scoring the speaking sections. The immediate goal of this effort was to improve the scoring efficiency of the TOEFL iBT Speaking Practice test while maintaining quality comparable to that of trained human raters. The long-term goal was to provide instructional and diagnostic feedback based on automated features in addition to providing valid and reliable total test scores. The result of this effort was the release of SpeechRater v1.0 for use in the TPO in November 2006.

The TOEFL iBT Speaking Practice test, like the TOEFL iBT Speaking Section, contains six tasks. The first two are independent tasks that ask candidates to speak about familiar topics based on their personal experience or background knowledge. The purpose of independent tasks is to measure the speaking ability of examinees independent of their ability to read English or comprehend spoken English. The remaining four are integrated tasks that engage reading, listening and speaking skills in combination to measure the communication skills typically required in campus-based situations and in academic courses. The entire test takes approximately 20 minutes. For each of the six tasks, the examinees are allowed a short time to prepare their response and then 45 to 60 seconds (the time limit varies by task type) to provide their response in a spontaneous manner.

The scoring rubric used by human raters to evaluate the responses to the TOEFL iBT Speaking Practice test is identical to that used for the TOEFL iBT Speaking Section. The raters issue a holistic score for each response on a score scale from 1 to 4 that is based on three key categories of performance: Delivery, Language Use, and Topic Development (see Xi & Mollaun, 2006 for the scoring rubrics).

A BRIEF OVERVIEW OF SPEECHRATER V1.0

SpeechRater v1.0 provides instant score feedback for the TOEFL iBT Speaking Practice test. It consists of three major components: the speech recognizer and feature generation programs, the scoring model, and the user interface. The speech recognizer and the feature generation programs are closely interrelated and can be considered as one integrated component that generates the scoring features. The speech recognizer decodes the input audio files into recognized words and utterances; then the feature generation programs extract the scoring features indicating different aspects of speaking performance, based on various output that the speech recognizer produces, which may include words uttered, pauses, pitch, energy, etc. The second component is the scoring model that scores responses to individual tasks based on the scoring features and summarizes the scores across multiple tasks. The last component is the user interface that provides the users with the score report and advisory information about how to interpret and use the scores. Details about different components of this system are not included in this paper, but interested readers could refer to Xi et al. (forthcoming) for more information.

AN ARGUMENT-BASED APPROACH TO VALIDATING SPEECHRATER V1.0

This section illustrates the application of the argument-based approach to validating SpeechRater v1.0. As Clauser et al. noted, the use of automated scoring will not only impact the strength of the evaluation inference, which links test performance to observed test scores, but also the subsequent inferences in the validity argument. This is described as the "ripple effects" of automated scoring that "extend through each step in the argument" in Clauser et al. (2002). To position automated scoring in an interpretive validity argument for using the test scores for a particular purpose, a general description of the chain of inferences resulting in a decision based on language test scores is provided below. (For an elaboration on building a validity argument for a language test, see Chapelle, Enright, & Jamieson, 2008).

Figure 1 illustrates the mechanism under which various types of inferences can be organized conceptually to link a sample of test performance to score-based interpretations and uses. The process of establishing an inferential link involves building an informal argument. In particular, each inferential link rests on certain assumptions that need to be backed by evidence.

Each inference, if sustained, becomes the grounds for the subsequent inference in the argument. The first link from a sample of language test performance to test scores hinges on the assumption that performance on a language test is obtained and scored appropriately to yield accurate scores for the intended use (Evaluation). The second link is from an observed score to a universe or true score. The pertinent assumption is that performance on language test tasks is generalizable over similar language tasks in the universe, raters, test forms and occasions (Generalization). In order to support this link, evidence is needed that the errors incurred in the measurement process are minimized to a level where we can be sure that if a test taker were given similar language tasks, rated by different raters, or administered in an alternate form or the same test on a different occasion, he/she would receive similar scores.

The third link between a universe score and an interpretation is crucial in the overall validity argument, because it bears on whether test takers' performance on the test provides adequate evidence about their language abilities that underlie their language performance in a target domain beyond the test. The assumptions are that test scores reflect the quality of language performance on relevant tasks in the real world (Extrapolation) and that speaking abilities and processes revealed by language test tasks vary in ways that are consistent with models of communicative competence in academic contexts (Explanation). At this link, meaning can be attached to the universe score in two potential ways to support valid interpretations of the assessment results. The universe score can be interpreted by drawing on a theoretical construct (e.g. a communicative competence model) that underlies consistencies in test takers' performances. For assessments for which specific domains of generalization can be defined, this representation of the meaning of assessment results is further contextualized in the domain to which the test scores are intended to be generalized. In some instances, in the absence of a strong construct theory, the extrapolation of test performance to the intended domain may sustain the link from the universe score to the score interpretation. The fourth link, utilization, connects score-based interpretations and decisions. The assumptions are that the test scores and other related information provided to users are *relevant, useful,* and *sufficient* for making intended decisions and that they promote positive effects on teaching and learning (Utilization) (Bachman, 2005).

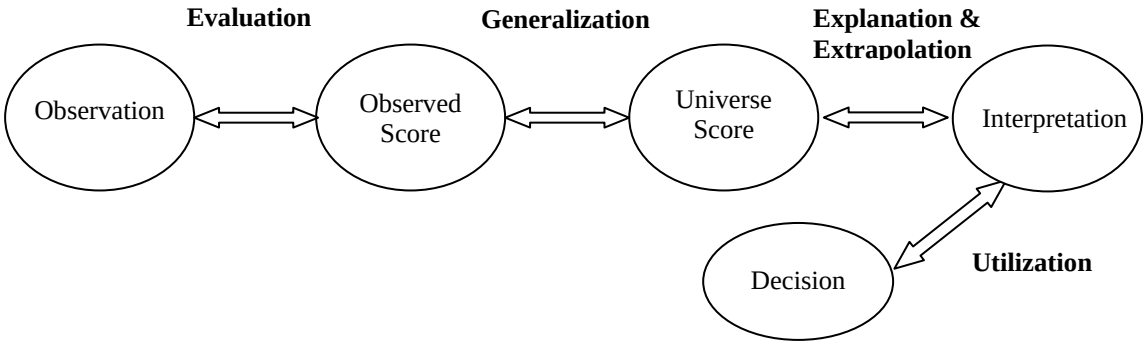


Figure 1. Links in an interpretative validity argument
(Modified after Kane, Crooks & Cohen, 1999 and Bachman 2005)

In summary, in the process of building a clear and coherent chain of reasoning, more and more meaning is attached to a sample of test performance and the corresponding score to justify the final score interpretation and use. These different meanings are tied to having accurate scores, generalizable scores, meaningful scores, scores that indicate domain performance, scores that are useful for decision-making, and scores that have beneficial consequences.

When automated scoring is integrated into an assessment, its most immediate effects seem to be on the accuracy of the resulting scores, thus pertaining to the Evaluation inference. This is also the aspect of automated scoring that has been most heavily researched. However, the effects of automated scoring may extend beyond this and be evident through all of the subsequent inferences. At each stage described above, automated scoring may introduce enhancements to validity that human scoring may not be able to offer or pose threats to validity in ways that are not typical of human scoring.

The use of automated scoring can potentially enhance the validity argument that supports the intended use of test scores. Specifically, it allows the designer of an automated scoring system to maximize construct representation by selecting construct-relevant response features and combining them to produce scores in a way that best represents the construct (Bennett & Bejar, 1998), thus contributing to the strength of the Explanation inference. This degree of control is not possible with human scoring. In addition, an automated scoring system applies the defined rating criteria consistently. It can thus improve score generalizability and strengthen the Generalization inference by eliminating differences in rater leniency or harshness, in raters' judgments over tasks, occasions or combinations of them.

Nevertheless, what may come with the systematic control of the construct is systematic error due to construct under-representation or construct-irrelevance. This is particularly true for a scoring system for a construct as complex and challenging as speaking proficiency. Conceptualizing and implementing speech features that indicate the key criteria human raters use to score spoken responses presents immense challenges. The tendency to extract easily quantifiable aspects of the performance due to the limitations of current speech technologies would potentially result in construct-irrelevant features or features that do not represent the full construct. In addition, given the complexity of human raters' decision-making processes involved in rating speaking, it obviously is not an easy task to design a scoring system that adequately reflects those processes. Even a scoring solution informed by expert judgments may not be adequate in representing the intended construct, depending on the qualifications of the experts and the rigor with which the work is conducted.

The systematic error introduced by automated scoring may impact more than one of the inferences that lead to the interpretation and the use of the scores. For example, automated scoring may reduce task specificity by disproportionately capturing aspects of speech that are relatively stable across tasks, thus improving the score generalizability and strengthening the Generalization inference. However, it may have reduced the task

specificity in undesirable ways. It may compromise the explanatory power of the scores in representing the constructs by failing to include some aspects of speech that are construct-relevant but are less stable across tasks, thus weakening the Explanation inference.

Replacing human scoring with automated scoring may also change users' perceptions of the assessment and the way they interact with the assessment tasks. Knowing that scores are produced by an automated system rather than human raters, users may also interpret and use the scores differently than they would use scores from human raters. Further, users typically perceive automated scoring as inferior to human scoring, although the latter is also prone to error. Their perceptions are sometimes misguided by some general misperceptions about automated scoring and may not be motivated by the specifics of a particular scoring system. Therefore, it is important to investigate how automated scoring may impact the Utilization inference through investigations of the aspects discussed above.

Since a validity argument is only as strong as its weakest link (Kane, 1992), it is critical to identify all the potential threats to the various inferences and provide counter-evidence against the rebuttals. The validation efforts should focus on providing counter-evidence that discounts these rebuttals.

To build and evaluate a validity argument for SpeechRater v1.0, four basic steps are involved:

1) Clearly state the intended interpretation and use of the automated scores on TOEFL iBT Speaking Practice test;

2) Articulate the network of inferences that lead to the intended interpretation and use and the associated assumptions that will lend support to each inference if backed by evidence;

3) Identify critical rebuttals that may weaken each inference as a result of using automated scoring; and

4) Collect and integrate evidence to reject the potential rebuttals associated with each inference.

The first three steps will yield an interpretive argument, the plausibility of which will then be evaluated in Step 4 in the context of a validity argument. This paper will address the first three steps and demonstrate the process of developing an interpretative argument.

The goal of developing SpeechRater v1.0 was to support the intended use of the product, i.e., help students better prepare for the TOEFL iBT Speaking and gauge their own readiness to take the official test. The claim we intend to support is:

> The SpeechRater v1.0 score is a prediction of the score on the TOEFL iBT Speaking Practice test a test taker would have obtained from trained human raters. The entire practice experience

> can help familiarize test takers with the content and format of the TOEFL iBT Speaking test so that they can better prepare for it. This score can be used by the test takers to help them self-evaluate their readiness to take the TOEFL iBT Speaking test.

This claim clearly specifies the intended low-stakes use of the TOEFL iBT Speaking Practice test and the score that SpeechRater v1.0 produces. Although this claim states what the SpeechRater v1.0 intends to do, it also conveys, although not explicitly, what it does not do. First, it does not intend to predict a candidate's potential performance on the TOEFL iBT Speaking test, which is taken under operational testing conditions. The motivation and anxiety levels of the candidates may be different when taking the official test versus the practice test. When taking the real test, candidates may be more motivated but more nervous. In addition, candidates can make several attempts on each task in the practice test whereas they are allowed only one attempt on each task in the official test. When taking the practice test, candidates could also choose to use more time to plan a response before starting to record it, but this option is not available for the official test. However, a candidate may be able to self-evaluate his/her readiness for the official test, knowing the conditions under which he/she has taken the practice test. A candidate could potentially choose to take the practice test under the timed mode and make his/her best effort to respond to each task as if he/she were taking the official test. Only under these circumstances would a candidate be able to assess his/her own readiness to take the official test.

Second, SpeechRater v1.0 does not intend to explain why a candidate receives a certain score. More specifically, the scoring model of the SpeechRater v1.0 does not mimic exactly how a human rater would have scored a test. It only intends to use meaningful speech features that indicate different aspects of candidates' speaking performance to predict the score of a human rater.

Further, SpeechRater v1.0 does not provide diagnostic feedback, although this is a long-term goal. It provides only a single score without any detail about why the score was ontained.

Table 1 shows the most common types of inferences that need to be verified to support the claims we would like to make based on scores generated by the SpeechRater v1.0. The crucial rebuttals that may potentially undermine the validity of these claims are also stated, associated with the inference to which each pertains. Failures to provide evidence to reject any of these rebuttals related to the critical inferences would potentially weaken the entire argument.

Guided by this framework, different lines of evidence can be organized into these five areas and synthesized to evaluate the soundness of the validity argument. Summarized below is key evidence relevant to each area that can potentially be gathered.

Evaluation. The relevant evidence includes the association between human and SpeechRater scores indicated by various well-established measures such as correlation and kappa. Different scoring methodologies that are used to produce SpeechRater scores

Table 1 Areas of emphasis for validity of SpeechRater v1.0 and associated rebuttals

Inference	Assumptions	Rebuttals
Evaluation	Automated scoring results in scores that accurately represent the quality of the performance on the practice test.	1. The scoring algorithm under- or misrepresents the construct or introduces construct-irrelevance so that the resulting scores are not accurate.
Generalization	The scoring model can generalize to new tasks and samples of candidates and the automated scores are generalizable over tasks.	1. The scoring model is built from insufficient or unrepresentative samples. 2. The scoring model does not generalize to new tasks or independent candidate samples. 3. The automated scores do not generalize across tasks.
Extrapolation	The automated scores reflect the quality of performance on relevant real-world speaking tasks in an academic environment.	1. Candidates' automated scores are not related to their levels of performance on real-world speaking tasks in an academic environment.
Explanation	The automated scoring model captures aspects of speaking performance in a manner that is consistent with theoretical predictions about speaking abilities used in an academic setting.	1. The automated scores are not adequate in *explaining* examinee performance in the domain. 2. The speech features used in scoring models are not well-linked to the rubric, introducing construct-irrelevance. 3. The speech features do not cover the key criteria defined in the rubric very well, resulting in construct under-representation. 4. The speech features are not combined in a meaningful way to produce scores. 5. The scoring model disproportionately captures aspects of the rubric that generalize across tasks, reducing task specificity in an undesirable way so that the construct is under-represented.
Utilization	The automated test scores and other related information provided to candidates are *relevant*, *useful*, and *sufficient* for them to make intended decisions and promote positive effects on teaching and learning.	1. The predicted scores and other information communicated to the candidates do not provide relevant, useful and sufficient information for them to gauge their readiness to take the TOEFL iBT Speaking test. 2. The automated scores negatively impact users' perceptions of the assessment and the way they interpret and use the scores. 3. The automated scoring system does not promote positive washback effects on English language teaching and learning. 4. Other potential negative consequences of SpeechRater v1.0 are not anticipated or minimized.

based on automatically extracted speech features can be compared based on the strength of the association between the model predicted scores and the human scores. The soundness of the statistical principles underlying each methodology is also an important consideration in employing a particular scoring methodology.

Generalization. This inference draws support from two types of evidence. One concerns the procedures for developing and evaluating the scoring models such as the adequacy of the sample size, representativeness of the sample, and absence of overlap in speakers between the scoring model training and evaluation data. The other type of evidence includes the generalizability of the SpeechRater scores across different tasks that can be estimated using Generalizability studies (Cronbach, Nageswari, & Gleser, 1963) or other established methodologies. The score generalizability estimates could be compared to those obtained for human scores and typical figures acceptable for a practice context.

Extrapolation. The potential evidence supporting this inference is the association between SpeechRater scores and scores on criterion measures of students' academic speaking ability, such as faculty or English instructors' ratings of their students' speaking proficiency.

Explanation. Evidence supporting this inference is conceptual and judgmental in essence. In particular, two essential qualities of the SpeechRater scoring model need to be verified to argue that it captures aspects of performance in a manner that is consistent with theoretical predictions about speaking abilities used in an academic setting: the construct relevance and coverage of the features and the defensibility of the way they are combined. The evidence involves largely judgments of these qualities by experts who have an intimate understanding of the construct the assessment is designed to measure, the conceptual meaning of each scoring feature used, and the way the scoring features are combined through a statistical model to produce a score that indicates the overall quality of performance.

Utilization. Arguments for the usefulness of the SpeechRater v1.0 scores for self-evaluations of readiness to take the official test are supported by an analysis of the magnitude of the prediction error in relation to the intended score-based decision. Arguments about potential consequences of the SpeechRater v1.0 can be made based on the score report, and the advisory information communicated to the user about the limitations of the system and the intended use of the scores can be included as part of the user interface. Additional evidence may include investigations of user perceptions, e.g., to what extent the awareness of the scores being produced by a machine impacts the way a user interprets and uses the scores, as well as the impact of using automated scoring on teaching and learning practices.

The aspects of the validity argument that require full support are dictated by the intended use of the assessment scores. For example, the areas of emphasis for validating an automated scoring system intended for a practice environment may differ from those for a system employed in an assessment for high-stakes decisions. If automated scores are

intended to support high-stakes decisions, all the five inferences discussed above need to be fully supported. The Explanation inference is especially important—if the automated scoring model under- or misrepresents the construct of interest, test takers may be misguided to focus on the wrong things or omit important things in their test preparation. It may also make the assessment more vulnerable to new types of cheating and test-taking strategies that would negatively impact the trustworthiness of the scores. An automated scoring system that under- or misrepresents the construct may also incur negative washback effects on teaching and learning and hurt the credibility of the test program.

Given that this initial version of SpeechRater focuses on providing prediction of human scores at a level acceptable for low-stakes decisions in practice environments rather than diagnostic feedback on learners' strengths and weaknesses in speaking, three of the five inferences particularly need adequate backing by relevant empirical or judgmental evidence: Evaluation, Generalization and Utilization. The Evaluation inference pertains to the accuracy of the automated scores; the Generalization inference concerns the stability of the scoring model and the generalizability of the scores across different tasks; and the Utilization inference is related to the sufficiency, relevance and usefulness of the score and other related information provided to candidates for making self-evaluations of their speaking performance. Although the Extrapolation and the Explanation inferences are important, adding meaning and value to the SpeechRater scores to support the subsequent Utilization inference, it is less critical for them to be fully supported for this version of SpeechRater.

Based on the validation framework discussed above, the relevant evidence pertaining to each inference can be integrated and evaluated. Then the overall strength of the validity argument can be evaluated in light of the critical inferences that need adequate backing to support the intended claims of this version of SpeechRater.

CONCLUSION

This argument-based approach to validating an automated scoring system drives researchers to consider in a systematic way what and how much evidence is needed to justify the use of an automated scoring system in an assessment for a particular purpose. This principled approach can guide us to think through the process of articulating an interpretative argument for using an automated scoring system as well as collecting and evaluating evidence to support a validity argument.

REFERENCES

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater v.2.0* (Educational Testing Service Research Report No. 04-45). Princeton, NJ: Educational Testing Service.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment* Quarterly, *2*(1), 1-34.

Chapelle, C. A., Enright, M. K., Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*™. Mahwah, NJ: Lawrence Erlbaum.

Clauser, B. E., Kane, M., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413-432.

Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology, 16,* 137-163.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112* (3), 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-35.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.18-64). Washington, DC: American Council on Education/Praeger.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST) (TOEFL iBT Research Report No. 1). Princeton, NJ: Educational Testing Service.

Xi, X., Higgins, D., Zechner, K., Williamson, D. M. (forthcoming). *Automated scoring of spontaneous speech using SpeechRater v1.0*. ETS Research Report Series.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391-412.

Part III

Learner Data, Diagnosis, and Language Acquisition

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

A Framework for Cognitive Diagnostic Assessment

Eunice Eunhee Jang
Ontario Institute for Studies in Education
of the University of Toronto

Researchers and teachers would like to improve assessments so that they can be used diagnostically to evaluate and monitor learners on particular aspects of their language skills. Current technical knowledge in assessment, however, is better suited to discriminating among learners by locating them on a continuous unidimensional scale. This paper discusses an approach to assessment intended to provide a much finer grained representation, cognitive diagnostic assessment (CDA), which is intended to inform learners of their cognitive strengths and weaknesses in assessed skills. The basic premises of CDA are promising, but this approach is not currently used in operational language assessments and further exploration is necessary to achieve this goal. As a first step, I explain the main tenets of CDA, and summarize my research using CDA for the assessment of ESL reading skills. Based on this research, I suggest a framework aimed at implementing CDA in practice by integrating it into computer-assisted language learning environments. The framework includes the use of diagnostic feedback by educator clientele.

INTRODUCTION

Proficiency and achievement testing has been criticized for its limited representation of knowledge and learning process (Glaser, 1994; Linn, 1990) and for its lack of diagnostic information to inform students of their strengths and weaknesses in a specific academic domain. As standardized tests are increasingly recognized to be unsatisfactory for guiding learning and evaluating students' progress (Mislevy, Almond, & Lukas, 2004), testing communities call for more diagnostic test information that allows for meaningful interpretations and the fair use of test results for improving instructional design and for guiding students' learning. Technical knowledge in assessment is, however, much less developed in this area. As a consequence, the guidance that language testing specialists take from educational measurement for proficiency testing is not available for operationalizing diagnostic language assessments.

In this paper, I introduce cognitive diagnostic assessment (CDA), an approach for design and interpretation of diagnostic assessment which appears to hold promise for language assessment. I explain *presuppositions* underlying the CDA framework and address *conditions* for valid CDA applications. I summarize my experience in applying CDA in an

empirical study (Jang, 2005) of L2 reading comprehension assessment from the *LanguEdge* courseware (ETS, 2002). Finally, based on the outcomes of this research, I describe a framework for optimizing CDA by integrating it into computer-assisted language learning environments.

DIAGNOSTIC ASSESSMENT FOR SECOND LANGUAGE LEARNERS

Dissatisfied with the prevalence of proficiency testing, language testing researchers increasingly call for more descriptive test information and detailed score reporting for improving instructional designs and guiding students' learning (Alderson, 2005; Bailey, 1999; Shohamy, 1992; Spolsky, 1990). Despite the necessity of research into the diagnostic score reporting processes, few empirical studies have examined the use of diagnostic reports in the context of teaching and learning. The use of diagnostic feedback needs to be understood by considering different beliefs about learning and different pedagogical approaches that teachers and educators hold. Equally important are learners because diagnostic feedback may have different effects depending on the learners' competency levels, cognitive and metacognitive learning styles, or learning context (Kunnan & Jang, forthcoming). Diagnostic feedback needs to be descriptive and interpretable so that it can help learners to take actions to close the gap between their current competency level and their desired learning goals (Black & Wiliam, 1998).

Despite its apparent utility for these purposes, diagnostic assessment has not received much attention, compared to proficiency and achievement testing, and therefore, few diagnostic assessment instruments are available for teachers to use in classrooms (Alderson, 2005). Research is needed to develop diagnostic testing that includes cognitive tasks suited for diagnosing learners' strengths and weaknesses in the tested skills. Such diagnostic tests need to be based on a systematic design framework that involves multiple steps (Davidson & Lynch, 2002; Mislevy, Steinberg, & Almond, 2003; Pellegrino, Chudowsky, & Glaser, 2001). The design framework of CDA, which appears to offer a useful perspective, can include: (1) defining the learning and instructional goals that serve as criteria for the content of diagnosis; (2) designing specific tasks that are diagnostically informative in evaluating a learner's competency in light of the learning goals; (3) developing a scoring system that allows for fine-grained diagnostic information; and (4) optimizing the reporting of diagnosis to maximize its use as intended. Such a design framework is obviously of interest for development of diagnostic assessment of language ability and worthy of exploration.

COGNITIVE DIAGNOSTIC ASSESSMENT

Cognitive diagnostic assessment (CDA) is a relatively new diagnostic assessment approach that is aimed at providing formative diagnostic feedback through a fine-grained reporting of learners' skill mastery profiles (DiBello, Roussos, & Stout, 2007; Embretson, 1991, 1998; Hartz, 2002; Nichols, Chipman, & Brennan, 1995; Tatsuoka, 1983). The CDA approach combines theories of cognition with statistical models to make inferences about

learners' mastery status for the tested skills. The cognitive skills or attributes refer to processes and strategies that test takers utilize to correctly solve tasks. Cognitive skill profiles summarize a learner's competencies in the tested skills.

Assessment of Learning vs. Assessment for Learning

When the purpose of an assessment is to evaluate and monitor learners on particular aspects of skills, a fine-grained representation of the competencies is necessary. CDA is intended to guide test developers to develop such a representation to be used for informing learners of their cognitive strengths and weaknesses in assessed skills. This assessment purpose contrasts with that of an assessment intended to discriminate among learners by locating them on a continuous ability scale. In this case, a unidimensional representation of learners' competencies in the subject domain should suffice.

CDA is aimed to serve as assessment used *for* learning and *as* learning process rather than assessment *of l*earning outcomes. The perspective of assessment *of* learning views assessment as a tool for summative evaluation of how much of the curricular goals the students achieved and how prepared they are to move to the next level in education (i.e., grade promotion, graduation, certifications). In this case, assessment results are used to make inferences about an individual test taker's general language ability with reference to other test takers in the normative group. Aggregated test scores based on unidimensional scaling are commonly reported even though the construct of language competency is often operationalized into a set of discrete skills. Reliability and accuracy in discriminating among individuals become the primary concerns. In such testing scenarios, the relationship between the assessor and the test takers is unidirectional and hierarchical.

The CDA approach is intended to promote assessment *for* learning by providing teachers with information needed to modify instruction and learning in classrooms. Teachers can use the formative diagnostic information to redesign instructional approaches, evaluate instructional resources, and remediate students' weaknesses. The CDA approach can promote the students' engagement in learning by encouraging them to use assessment as a learning tool. As critical assessors of their own learning, students are actively engaged in various learning and assessment activities by making sense of information, relating it to their prior knowledge and experience, and using it for planning new learning.

Specification of Cognitive Skills

The CDA framework guides diagnosis by bringing together cognitive science and psychometrics to make substantive assumptions about the processes and knowledge structures that a learner would use in completing tasks. CDA requires the test be based upon a substantive theory of the construct that describes the cognitive processes through which a learner performs on tasks and, at the same time, it requires clear specifications that delineate item or task characteristics that are intended to elicit the cognitive processes (Embretson, 1998). These prerequisites present a challenge for test designers.

Definitions of how learning takes place and how a competency of interest can best be assessed vary depending on the contemporary theories of learning (Greeno, Collins, & Resnick, 1996; Resnick & Resnick, 1992). The CDA approach is grounded in the cognitive view of learning and knowledge acquisition in which knowing is presupposed as a systematic processing of information. In this cognitive view, learners understand concepts through reasoning, they use cognitive and metacognitive strategies for problems, and in turn they transfer new knowledge to other tasks. In contrast, a socio-historic view of learning presupposes that learning takes place through participation in socially organized practices such as formulating and evaluating questions critically, making inferences based on prior experience, and presenting reasoned explanations. These are important differences that need to be addressed in the CDA design because the use of the CDA is greatly influenced by different views of learning that teachers and educators across educational and cross-cultural contexts may have. Therefore, theories of cognition and learning should be clearly defined and examined within the context of the subject domain when considering the CDA application. This obviously presents a challenge in second language acquisition, where multiple theoretical perspectives are brought to bear on the explanation of learning a second language.

What is needed to begin to explore these issues is empirical research applying CDA to second language assessment. Two approaches can be taken to move forward on this program of research; (1) an inductive approach to creating a set of diagnostic items or tasks that allow us to infer the skill processes and knowledge structures of interest; and (2) a retrofitted approach (or reverse-engineered approach) to extracting cognitive processes and skills from existing tests in hope of obtaining richer information than what unidimensional scaling can offer. The first approach requires that the cognitive skills of interest should be explicitly targeted during item and test development. All relevant skills should be considered with an appropriate balance of cognitive skills. Documentation of the process of item and test development is required to enhance the transparency of targeted cognitive skills for the CDA users. The second approach has developed in part due to a lack of existing diagnostic tests. CDA approaches have been applied to existing achievement or proficiency tests in hope of providing fine-grained diagnostic feedback beyond what aggregated test scores can offer (See, Buck & Tatsuoka, 1998; Sheehan, 1997; Tatsuoka, 1990). Such detailed information about the important skills required for success on a high-stakes test should be valuable to examinees and at the same time it should provide an opportunity to better understand the use of CDA.

AN INVESTIGATION USING CDA

I conducted a study using the second approach to explore the utility of CDA as a framework for generating and reporting diagnostic information about reading to ESL learners. The reading assessment that I used was *LanguEdge*, which was developed to assist teachers and learners with preparation for the high-stakes TOEFL iBT, thereby serving as an instructional tool for use in the English as a Second Language (ESL) classroom. However, like on TOEFL iBT, scores were reported at the level of the section

(e.g., reading) and the total score. The research sought to provide a finer granulation of useful diagnostic reporting through the use of CDA and to assess the value of this information for test takers.

Research Design

The study consisted of three phases. In the first phase, 11 students took the *LanguEdge* assessment to provide an evaluation of their reading skills and participated in think-aloud verbal protocol analyses during test-taking. The items from the reading assessment were content-analyzed. Student performance data from the field test (N=2770) were analyzed statistically. Using these identified skills, the characteristics of skill profiles estimated by the Fusion Model (Hartz, 2002) were examined in the second phase. The third phase involved 28 ESL students and two teachers recruited from two TOEFL preparation courses. The students took pre- and post-instruction diagnostic tests made of items from the *LanguEdge* reading comprehension tests and completed self-assessment questionnaires. They received individualized diagnosis report cards at both junctures. Using interviews, classroom observations, and surveys, the usefulness of the diagnostic feedback was evaluated. The analysis investigated the use of CDA from many perspectives (see Jang, 2005), but here I summarize evidence concerning the validity of the inferences made on the basis of the CDA analysis and the validity of the use of the resulting diagnostic information.

Validity of Inferences

CDA statistical models are developed with a strong assumption about how cognitive skills or combinations of the skills influence students' test performance. In other words, the statistical model assumes particular types of inferences can be made on the basis of observed performance. To support inferences, evidence is needed pertaining to the nature of cognitive skills, ways in which such skills are involved in problem-solving processes, and the extent to which such skills are identifiable independently. The data obtained from the think aloud protocol were used to evaluate the claim that cognitive skills included in diagnosis profiling reflected the kinds of strategies and processes that learners used when solving the assessment tasks.

The analyses of the think-aloud verbal data indicated that students' processing of reading skills was concurrent and interactive. By concurrent, I mean that the students utilized multiple strategies simultaneously to process the textual information. In many cases, different strategies led to success in solving the same question for different students. For example, processing of vocabulary knowledge varied to a great extent depending on the students' language background, prior knowledge, and the degree of the context dependency. By interactive, I mean that the students tried to gather information resources from various sources such as the text, questions, and their prior experience and knowledge. The interactive processing was observed more often with tasks involving textually implicit information, inferring, or determining word meanings. They actively negotiated textual meanings by utilizing resources at the different levels of processing.

The observed evidence of reading processes posed some challenges for the goal of identifying the cognitive skills for the CDA approach. First of all, when students utilized multiple, yet different strategies to solve a task, it was challenging to determine which strategy would be essential for success in the task. Secondly, a close examination of the reported reading processing strategies indicated that the students' reading processes were sensitive to various organizational patterns of the texts. For example, when asked to choose the options that best summarize the main idea that appeared in the descriptive/expository textual type, the students tended to recall textual information without recourse to the text. When reading a text written in a more complex rhetorical structure, the students appeared to rely more on the text to confirm contrasting ideas by locating them in the text. This suggests that the choice of a textual type may influence the reading processes and the strategy types. The implication of this observation for the CDA is that it is essential to ensure that a set of reading skills is specified carefully after taking into account such interactive relationships with various textual variables.

Overall, the reading skills identified from the analysis of the think-aloud verbal protocol data were consistent with the skill specifications of test developers. The reading skills were verified by five content experts' ratings with reasonably acceptable agreement rates. The results from the think-aloud data suggested a broader range of skills and strategies, and they supported fine-grained representation of the construct for the CDA application.

A second approach to assessing the validity of the inferences made from the CDA was to examine relationships between the CDA results and students' self-assessments. Results indicated that test takers' self-assessed ratings on their reading skills were positively correlated with the model-estimated skill mastery profiles. In-depth analysis of individual cases also confirmed that positive relationship, suggesting that the students' self-assessment can provide useful information for evaluating students' skill mastery profiles and that the CDA results were supported by the self reports.

Validity of Diagnostic Feedback Use

The use of diagnostic feedback was directly examined in two TOEFL preparation courses by examining the perspectives of the students and the teachers. The students welcomed the skills diagnostic feedback provided in their report cards, called *DiagnOsis I* and *II*. Two teachers received summary diagnostic reports as well. The majority of the students found it very useful to understand their strengths and weaknesses in reading skills. Interviews with the students and surveys showed that roughly one half of the students confirmed the accuracy of the provided diagnostic information. Further, the students seemed to judge the accuracy of the skills diagnostic information by relating it to their own self-assessment on the skills. Students with poor skill profiles showed emotional frustration on the skills diagnosis results. They asked for more specific guidance for improving the weak skills.

Some concerns raised by the students are worth mentioning. One student's question about the meaning of a 'master' makes us think about what it means to be a master for a certain skill. This question is very important because diagnosing someone as a master or

non-master of a skill implies they should be recommended to take some future action. Students who received diagnosis with a large number of weak skills showed interest in how to improve them. They desired specific guides for the kinds of actions to take. As such, very good diagnosis results can also frustrate students because the actions they should take are less straightforward. As the student raised it, being a master for a certain skill obscures a future action. What course of actions can a master take as a result of diagnosis? It certainly does not mean that the student does not need to study any more. Thus, when calling someone a master, we need to be very clear about what is expected of a master. This implies that providing diagnostic information does not complete the act of diagnosis; it is only one step in a larger instructional context.

Another interesting aspect examined in the study was the extent to which students would agree on the linkage between the skills and the associated items. Only 18% of the students agreed that the example questions assessed the associated skills. The rest of the students expressed various alternative views as described in Jang (2005). One student pointed out the lack of sufficient test items for determining skill competency by stating that "I think the more questions I have, the more I can be convinced to know about my reading proficiency. But we don't have enough questions" (p. 172). Two students raised an issue about the extent to which the reported skills are independently divisible by stating "I think these questions assess the skills well, but I also think those skills can't be divided accurately because most questions need combined skills anyway (p. 172) " and "Actually I don't know how much these questions assess those skills correctly. If I could understand the whole passage well, it won't matter" (p. 172).

Although the students appraised the usefulness of the diagnostic feedback, the effect of the diagnostic feedback on learning remains uncertain. The study did not provide sufficient evidence to claim any direct effect of the diagnostic feedback on students' improved learning; especially because the study was conducted before the TOEFL iBT was launched. However, examination of changes of the students' skill mastery before and after the instruction revealed quite interesting patterns. The first pattern showed that the high-performing students' skill profiles exhibited stability over time while the second pattern was exhibited with a group of students who had improved significantly. The third pattern showed a group of students whose skill mastery was fluctuating and unstable. The observation of the different skill developmental patterns over time points to the importance of prolonged evaluation of learners' skill development, especially for the students with low-proficiency. In addition, fine-grained diagnosis can have the potential for providing the kind of information needed for such a longitudinal evaluation of skill trajectories.

Interviews with three teachers including the two teachers and one former teacher indicated that the teachers found the diagnostic feedback useful for raising students' awareness of their strengths and weaknesses in reading skills and for guiding their teaching. However, the teachers also raised some important issues concerning the use of diagnostic feedback. One male teacher pointed out that the use of diagnostic feedback may depend on the context of learning:

> We need to consider differences that lie between EAP (English for Academic Purpose) courses and test preparation courses that we are talking about now. In the test preparation courses, there might be more "teaching to the test" than in an EAP class. Such difference could be an important variable for evaluating the use of diagnostic feedback. (p. 176)

This indicates that the usefulness of skills diagnosis also depends on the purpose of learning and the context of learning. As such, diagnostic feedback may be beneficial for proficiency or achievement testing situations as long as there are low stakes attached to the test.

A female teacher raised her concern about a mismatch between the skills diagnostic approach and her own pedagogical beliefs:

> Knowing my students' strengths and weaknesses was very useful even though most of them needed to improve almost all skills after all. But I don't teach reading separately. I try to encourage students to study listening, reading, and structure simultaneously. So, I don't teach the reading skills included in this scoring report. (p. 174)

This implies that the use of the skills diagnosis does depend upon the degree to which it is compatible with the teachers' beliefs about teaching and learning.

Issues and Implications

The identification and use of diagnostic information was in some ways successful despite the process of retrofitting that was used to explore CDA in this study. However the retrofitting approach also presented some limitations.

The statistical evidence supported a relatively good fit of the model to the data, but a close examination of performance differences between masters and non-masters, as determined by the model-estimated skill mastery probabilities, indicated that 20 to 30% of the items failed to effectively discriminate masters from non-masters. A further examination of these items suggested that they exhibited extreme item difficulty levels. Although this result is not completely unexpected, it points to a significant problem associated with the use of non-diagnostic test for the CDA purpose. When the non-diagnostic test is developed for norm-referenced testing, the test includes items that adhere to the psychometric principle essential for creating a bell-shaped score distribution by including a wide range of item difficulty levels. Such a psychometric principle may not conform to the principle that guides diagnostic assessment.

The cognitive skills that formed the basis of the CDA were greatly constrained by the task types. When the test is not developed with diagnostic purposes in mind, the test may well include either too many or too few items for assessing particular skills. For example, while approximately 21% of the test items are vocabulary items for Form 1 of the LanguEdge RC test, the test does not have a sufficient number of items that elicit skills such as inferring authors' intention or summarizing the main ideas. Therefore, adequate specifications of cognitive skills become a quite challenging task when the CDA is applied to existing non-diagnostic tests.

Despite the limitations of the retrofitting approach to CDA, the research provided results that were interesting and useful to work with in order to envisage a broader framework for CDA. Such a framework would rely on the use of technology for implementation.

BENEFITS OF COMPUTER-ASSISTED CDA

To maximize the use of CDA for instructional practice and learning, diagnostic results from any assessments should be sufficiently aligned with the content of the curriculum. This is easier said than done because, even if a diagnostic assessment is developed inductively following the principled design framework, a generic diagnostic assessment in a traditional paper-and-pencil test format cannot address all of the curricular details specific to learning context. Technological integration is essential for realizing the potential of CDA.

Immediate Reporting of Diagnostic Feedback

The most important contribution of computer-assisted CDA might be provision of diagnostic feedback in a timely manner with no time lag. A significant delay between test administration and the reporting of test results, as carried out through the report cards in the research, is a key obstacle to teachers' use of the diagnostic information (Huff and Goodman, 2007). Integrating rigorous scoring or CDA calibration methods into the computer portal would also allow the CDA users to decide when and how to use diagnostic feedback. Diagnostic feedback does not have to wait until the end of the test administration. Instead, it can be provided in an interactive manner.

When performance-based assessment tasks are used for CDA, automated diagnosis and scoring systems can compensate for labor and time intensive scoring by human raters. Computer-generated diagnosis report cards can be immediately prepared for use by teachers. The teachers receive summary reports on their students' performance with detailed diagnostic information about areas that they need to improve. The students receive individualized diagnosis report cards for their review. The teachers and the students can have a conference to discuss the kinds of pedagogical actions that they need to take. School administrators and curriculum developers could also receive reports summarizing the students' strengths and weaknesses in tested skills. They could use the information to evaluate the effectiveness of the curriculum innovation.

Authentic Assessment of Learners' Skill Competencies

The computer-assisted CDA would help to overcome over reliance on traditional multiple choice item types by incorporating alternative item types that are designed to elicit cognitive skills and strategies in a more integrated manner. The computer-assisted CDA can provide authentic information about learners' skill competencies by using tasks designed for assessing integrated language skills such as: (1) summarizing orally or in writing after listening to a lecture; (2) simulating language use in context; (3) transforming information into a different form (tabulation, graphic representation); or (4) metacognitive

reasoning about the appropriateness of language use in a specific context. The majority of such authentic tasks would rely on natural language processing.

Utilizing Various Sources of Information for Diagnosis

The computer-assisted CDA allows the assessment developers and users to consider various sources of information beyond the correctness of the responses to test items. Learners' choice of distracters in multiple-choice items or response times retrieved in the computer database can provide diagnostically useful information about the learners' skill competencies. Current statistical advancements in various partial credit CDA models and scoring models utilizing information from the choice of distracters and response times would help expand the kinds of sources that can be used for diagnosis beyond the performance data.

In addition, the computer-assisted CDA can utilize information about non-cognitive aspects of learner characteristics in creating diagnostic skill profiles. Individual differences, such as socio-cultural and linguistic background and motivation, may need to be taken into account for the effective diagnosis of learners' skill competencies. Students' self-assessment of skill mastery and problem-solving strategies may enhance their metacognitive awareness of the effectiveness of strategy use and facilitate the use of diagnostic feedback for taking remedial actions to change their learning.

Flexibility for Customizing a Diagnostic Test

Computer-assisted CDA can provide a flexible interface that allows CDA users like teachers to customize the content of a diagnostic test by aligning it with what is taught throughout the instructional term. A sufficiently large item bank with a wide range of skills and task formats will allow teachers to design a diagnostic test that assesses specific skills in a manner similar to the instructional approach. Assessment developers can pilot test items to scrutinize their diagnostic power. For example, Diagnostic Information Indices (Jang, 2005) made available for teachers can inform them not only of item difficulty levels but also of the degree of diagnostic information of items being considered.

A FRAMEWORK FOR COMPUTER-ASSISTED CDA

The customized diagnostic assessment framework in computer-assisted environments entails many features. It emphasizes collaborations among various educational participants involved in testing and educational practice. The participants may include assessment specialists, teachers, students, school administrators, and educational policy makers. The collaborations need to take place throughout all of the phases of assessment development, implementation and evaluation. Figure 1 summarizes major collaborative activities of the computer-assisted CDA.

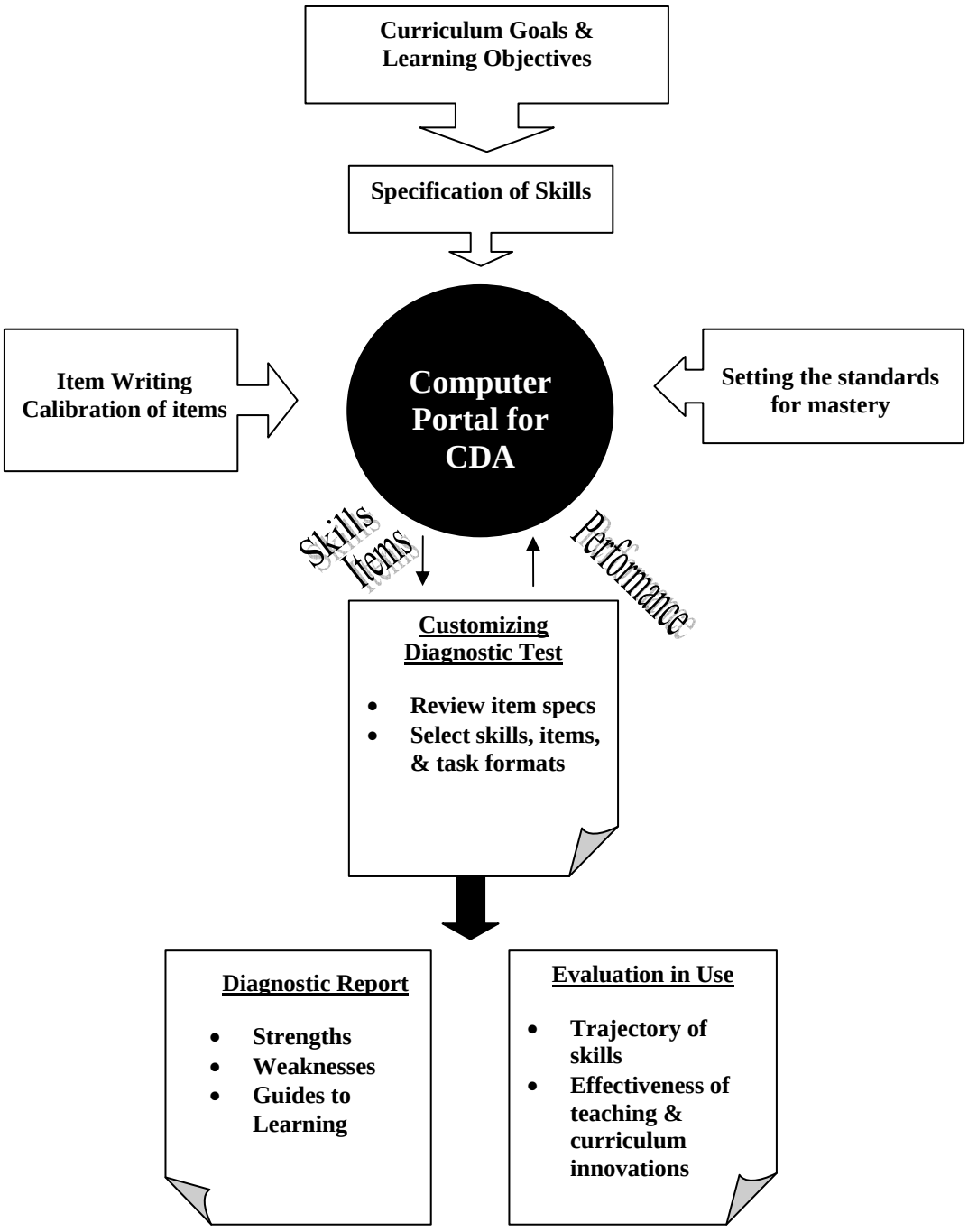


Figure 1. Customized computer-assisted CDA.

Curriculum goals and learning objectives

Development of the customized diagnostic assessment system starts by identifying curriculum goals and specifying learning objectives. Theories of learning are recognized and critically evaluated in light of the curriculum expectations and instructional approaches. Consistent with the curriculum goals, learning objectives are clearly specified by teachers. Teachers make explicit the target language skills, instructional activities to facilitate the development of the identified skills, and expected mastery levels of the skills throughout the school year. At this stage, although teachers are expected to play a primary role, assessment specialists and curriculum developers need to be involved in these activities in order to understand the curricular goals and learning objectives.

Item writing and calibration of items

Assessment specialists collaborate with teachers to develop item and skill specifications that delineate the diagnostic tasks that are intended to assess the skills included in the curricular and learning objectives. The skill specifications can be viewed more broadly than the item specifications in that they define cognitive processes and problem solving strategies in general terms. The item specifications provide the detailed information about task types, sample items, and the conditions for the administration procedure. Davidson and Lynch (2002) provide a comprehensive guideline for the collaborative item specification development process.

Once a sufficient number of items for the primary skills are developed, these items are pilot tested and calibrated through statistical cognitive diagnosis modeling. Information about each item's diagnostic capacity is prepared by assessment specialists, stored in the computer portal, and used as a resource for teachers when customizing diagnostic tests for their students.

Customizing the diagnostic test and setting the mastery levels

Here is a hypothetical scenario. Ms. Smith just completed a unit on critical evaluation of literature work for her ESL students. A pre-instructional diagnostic test indicated that her students were not competent in the following areas: (1) identifying authors' intentions; (2) understanding different styles of writing; (3) summarizing main ideas; and (4) critically evaluating literature in their own voices. After completing the unit, she wanted to know how much progress her students made in those areas. She entered the computer portal and reviewed the specifications of items associated with those skills. She carefully selected a set of items with varying item formats, difficulty levels and text types. She set an expected mastery level for each tested skill, which could serve as an initial parameter estimate ($P_{k's}$ in the case of the Fusion Modeling) in CDA modeling or as a cut-off point for determining skill mastery. When the students completed the test on the computer portal, the students' performance data were submitted automatically to the database for scoring. Computer-generated diagnosis report cards were immediately prepared for review by the teacher and her students. She met with individual students to discuss the skill profiles and ways to improve weak skills.

Use of diagnostic results

Diagnostic results from the aforementioned scenario can be used in many different ways. Teachers can use the results to reflect on their instructional methods and to plan remedial activities to help individual students. Teachers need to encourage students to be aware of their learning strategies and monitor the effectiveness of the strategies that the students use. The students can participate in various skill-development activities such as cued performance, modeling of higher-proficient students, or think-aloud verbal activities done either individually or in a small group. Individual differences, such as socio-cultural and linguistic background, and prior learning experiences, need to be taken into account for the effective skill-building activities. Formative diagnostic assessment needs to take place on a regular basis so that students' learning progress can be evaluated longitudinally. Teachers can use accumulated test results to communicate with parents, school administrators, and curriculum developers to enhance the quality of learning outcomes and allocate the necessary resources strategically.

FINAL REMARKS

Empirical evidence from the examination of retrofitted CDA approach to a L2 reading comprehension assessment prompted a revised CDA framework intended to guide use of CDA in second language assessment. The framework includes the provision for maximizing the use of the CDA for instructional practice and learning by aligning the content of the diagnostic feedback with the content of the curriculum. This alignment cannot be done using a traditional paper-and-pencil test due to various limitations. To overcome such limitations, I proposed that technological integration is essential for the potential of the CDA to be realized. Various advantages of the computer-assisted CDA framework were discussed. They include timely reporting of the diagnostic test results, the potential for using more innovative and authentic task formats, and the availability of the information about learners' cognitive and non-cognitive characteristics beyond their performance data for creating diagnostic skill profiles. Finally, I highlighted that the computer-assisted CDA allows for a flexible interface so that teachers can customize the diagnostic test to align it with instructional objectives and approaches. The research described in this paper sets the groundwork for continuing to explore the great potential for computer-assisted CDA approaches.

REFERENCES

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment.* London: Continuum.

Bailey, K. M. (1999). *Washback in language testing*. TOEFL Monograph Series Report. No. 15. Princeton, NJ: Educational Testing Service.

Black, P. J. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119-57.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven and London: Yale University Press.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitive diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 45-79). Vol. 26, Psychometrics. Elsevier Science B.V.: The Netherlands.

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 37*, 359-74.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-96.

Glaser, R. (1994). Instructional technology and the measurement of learning outcomes: Some questions. *Educational Measurement: Issues & Practice, 13*, 6-8.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-45). New York: Macmillan.

Hartz, S. M. (2002) A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). NY: Cambridge University Press.

Jang, E. E. (2005). A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Kunnan, A. J., & Jang, E. E. (forthcoming). Diagnostic feedback in language assessment. In Long, M., and Doughty, C. (Eds.), *Handbook of second and foreign language teaching.* Walden, MA: Wiley-Blackwell Publishers.

Linn, R.L. (1990). Diagnostic Testing. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 489-498). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to Evidence-Centered Design. CSE Technical Report 632, The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE). LA, CA: University of California, Los Angeles.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Nicols, P. D., Chipman, S.F., & Brennan, R. L. (Eds.). (1995). *Cognitively diagnostic assessment*. NJ: Lawrence Erlbaum Association, Publishers.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D.C.: National Academy Press.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative view of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333-352.

Shohamy, E. (1992). Beyond performance testing: A diagnostic feedback testing model for assessing foreign language learning. *Modern Language Journal, 76*, 513-521.

Spolsky, B. (1990). Social aspects of individual assessment. In J. de Jong & D. K. Stevenson (Eds.), *Individualizing the assessment of language abilities* (pp. 3-15). Avon: Multilingual Matters.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K. (1990). Toward an integration of Item Response Theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). New Jersey: Lawrence Erlbaum Associates Publishers.

Choo, J., & Kim, D.-H. (2008). Study on the analysis of learner data for the effectiveness of an ESL CALL program. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 132-148). Ames, IA: Iowa State University.

Study on the Analysis of Learner Data for the Effectiveness of an ESL CALL Program

Jinhee Choo
Doe-Hyung Kim
University of Illinois at Urbana-Champaign

Learner data can reveal important information about student performance and the effectiveness of a CALL program in terms of its tasks and feedback. In this paper, we demonstrate what sources of interaction from the CALL task interface can be collected to develop an informative student model. Student models can be an important contribution of CALL to SLA because both the process—the learner-computer interactions recorded—and the product—the progress learners make—may be useful for understanding how students learn a language through CALL instruction. We then discuss some quantitative statistical analyses from regression and categorical data analysis models from two separate empirical studies that were conducted with Korean ESL learners who worked on a CALL program developed to help ESL learners increase their awareness of consistent errors in academic writing. Although there was no significant linear relationship between time spent on the program and improvement between the pre- and post-tests, a marginal correlation between these two variables was found and other variables such as gender were related to performance and improvement of learners' language learning to a various degree. Furthermore, a survival analysis conducted with data from a particular task resulted in a model that described how students reacted differentially to three different feedback types. Examples from this application and other ongoing projects demonstrate a variety of informative process data collection methods within the task interface in addition to some limitations and implications of process data use.

INTRODUCTION

Computers can assist in the observation and analysis of student performance by allowing researchers to view video recording of learner interaction and analyze audio recordings. On the other hand, CALL programs in particular can be designed in such a way as to collect data that can yield insights about factors that affect learner performance on second language learning tasks. These data enable researchers to enhance the effectiveness of the CALL program in the areas of task and feedback design. In this paper we describe two studies of CALL learner performance. The first illustrates how data analysis could answer questions about whether various factors such as time spent on task, gender, and ESL language course experience were related to student learning. The second study is a follow-up of a previous study conducted by Kim (2005), in which the effectiveness of

types of feedback was analyzed using an advanced categorical data analysis method to confirm his findings. The results of these two studies bear on the findings of several other studies (i.e., Hegelheimer & Towers, 2004; Van der Linden, 1993) that examined which factors may affect learning via CALL programs and suggest ways in which additional features can be built into CALL programs to yield other behavioral data that can improve CALL programs and provide an effective measure of second language learning.

LEARNER DATA

Behavior tracking through screen-capture programs or process data maintained by the program itself allows researchers to examine the learning processes in detail (Beaudoin, 2004; Chapelle, 2003; Cowan, Choi, & Kim, 2003; Glendinning & Howard, 2003). Wible, et al. (2001) point out that programs that are capable of storing and tracking both teacher and student input can render the program more useful as information about the learner and teacher feedback gets accumulated into the database. Skehan (2003) argues for support software that will provide learners with pedagogic materials when the gaps in the learners' skills are detected, and for software that can provide a recorded indication of the learners' interlanguage development. Such arguments provide the rationale for creating a well-designed learner data collection to satisfy research needs and increase pedagogical effectiveness of CALL programs.

Approaches to Learner Data Analyses

There are various approaches to analyzing learner data. Performance information such as scores and time on task can be obvious indicators. Web-based programs connected to databases are not only able to store how frequently users click on tools or help features, but they can also keep track of how learners modify their output according to certain types of feedback.

Van der Linden (1993) examined learner data to investigate their preference for particular CALL feedback. Participants read prompts on the screen and responded by typing in sentences. They had unlimited attempts to answer each question, and had access to the answer at any time. It was hypothesized that the optimal learning condition would occur when students frequently accessed the feedback per item. Two significant behaviors were noted. High proficiency students utilized the optimal method of accessing the feedback until they figured out the correct answer. In contrast, lower performing students only accessed the answer and opted not to read the feedback. Thus, feedback preference seemed to be a good predictor of learner proficiency.

Whereas many CALL studies are conducted in a laboratory setting, Hegelheimer and Towers (2004) examined learner data from a study of the use of a CALL program, *New Dynamic English*, in an authentic environment with 94 female EFL students at a university in the United Arab Emirates for two months (using the program one day per week) to see if time spent on the CALL tasks and learner proficiency may influence the use and the effectiveness of various kinds of CALL options. Recorded learner data

included learner access to the microphone, headphone, repeat, speech recognition, ABC buttons (repeating and displaying the text simultaneously), and the glossary. The data revealed the overall usage pattern of the available options to the learners in addition to the total time interacting with the software, the placement test score indicating learner proficiency, and the shuffler level, which is a built-in adaptive testing mechanism. They found that although the use of the specific software features was widely variable among the learners, the use of certain options such as more frequent use of the ABC button by the lowest-performing group and more frequent use of the repeat button by the highest performing group could predict who the high performers and the low performers were. Time spent on the program was found not to be a significant predictor of success with a weak (t=0.347) and non-significant (p=0.730) correlation with student performance, though it showed some positive relationships with learner performance.

Data on Gender

According to Astleitner & Steinberg (2005), various gender differences have been identified within computer-assisted learning and whether they still exist within web-based learning remains an open question. For instance, there still remain some gender differences in the use of computer and attitudes towards computers among college students (Mitra, Lenzmeier, Steffensmeier, Avon, Qu, & Hazen, 2001), but recent research has found fewer overall gender differences in the frequency of computer use (Colley & Comber, 2003). As suggested by Astleitner & Steinberg (2005), females tend to use the learning modules more often than males when a course is about language education in web-based learning, though research results show that both gender groups were equal in terms of learning outcomes.

RESEARCH QUESTIONS

The research questions in this study were motivated by prior research suggesting that factors such as time spent on tasks, gender, and learner proficiency may have an effect on learner performance in a CALL program. Findings from previous research also indicated that both the interactions recorded on the computer and the learning outcomes should be integral to understanding how students learn a language through CALL. Reviewing the types of feedback students received and their subsequent corrections can reveal which type of feedback is most effective. In addition, various learner variables such as time, gender, and learner proficiency, may play a role on learner performance during the CALL instruction. However, there has not been much research on the effects of such variables in a web-based language learning environment. In this paper, we will address the following questions in this study via two quantitative analyses of learner data in CALL research using learner performance data collected from two previous experimental studies. The research questions are (a) whether there is relationship of time spent on the tasks, gender, and learner proficiency with learner performance; and (b) which feedback type results in more correct answers in a CALL program. For the first analysis, the specific factors that may influence the improvement of the learners' grammar skills while they use a CALL

program is going to be investigated through correlations and regression analyses. The second analysis examines the effectiveness of feedback utilizing an advanced categorical data analysis method to confirm the descriptive study conducted by Kim (2005).

MATERIALS AND METHODS

Materials (*The ESL Tutor*)

The CALL program used in this study was intended to help international students increase their grammatical accuracy in writing. The program was developed on the basis of error analyses of essays collected from students taking ESL courses at three different levels. The essays were used to develop a 221,556 word corpus of written English that represents writing from international students at a mid-western university in the United States. These students are placed into a class based on an English Placement Test (EPT) unless their TOEFL score is over 600 (or 250 in CBT). ESL 400, 401, and 402 courses are for graduate students, and most graduate students are placed into ESL 400 or 401 courses first and then move to the next level. The corpus encompassed ESL students' written products from these three classes in addition to classes from the Intensive English Institute, which prepares students to enter US universities—mostly undergraduate levels. Overall, the corpus has at least two (IEI vs. university students) and possibly three distinct proficiency levels (lower intermediate (IEI), intermediate (undergraduate ESL courses and graduate ESL courses such as 400) and advanced (graduate courses such as 401 and 402).

For the current study, we used performance data collected from *the ESL Tutor*, which was designed to provide explicit instruction on areas identified in the corpus as posing persistent grammatical errors for Korean students. The suggestions taken in the design of the program to improve grammatical skills can be summarized as follows: a) make key linguistic characteristics salient, b) offer modifications of linguistic input, c) provide opportunities for comprehensible output, d) provide opportunities for learners to notice their errors, e) provide opportunities for learners to correct their linguistic output, f) support modified interaction between the learner and the computer, and g) provide opportunities for the learner to act as a participant in L2 tasks (Chapelle, 2005). The structure of the program is shown in Figure 1.

In *the ESL Tutor*, students read through a lesson about a grammatical topic in Section A. They confirm what they learned through a grammaticality judgment task in Section B. Another lesson that compares and contrasts the learner's L1 sentences and equivalent English sentences is provided in Section C. Students finally engage in a highly interactive task of finding and correcting errors in Section D. Section E contains a longer passage with the same error types, and the Unit Test lets students test their knowledge by finding and correcting errors that cover multiple grammatical topics.

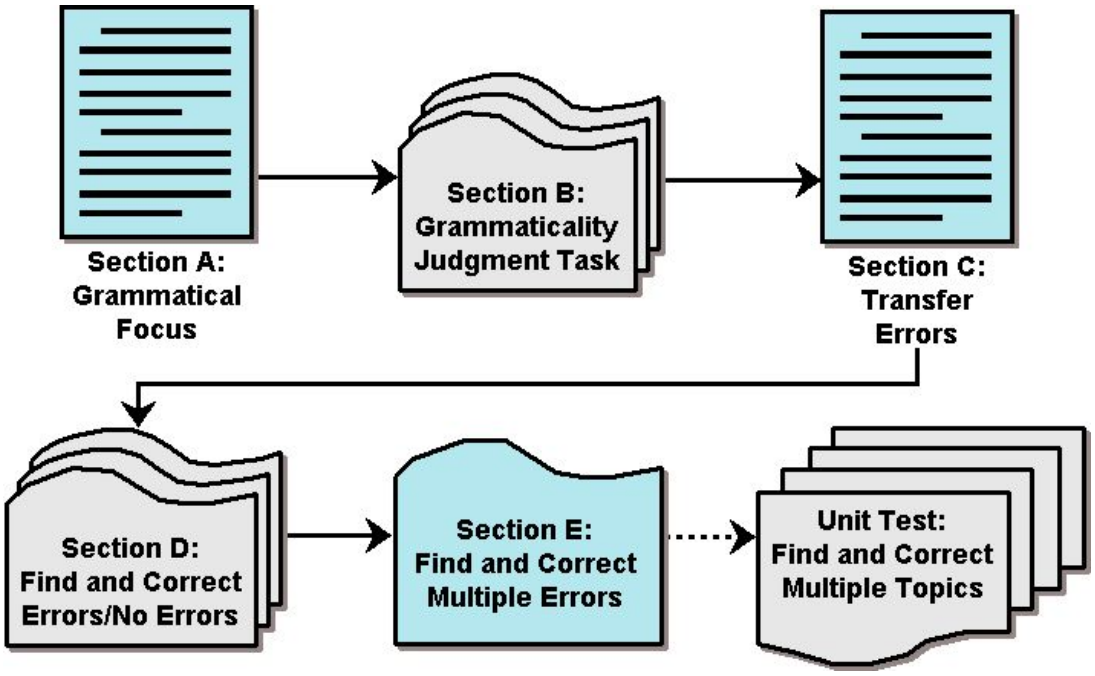


Figure 1. The program structure

The lessons provide explicit rules, models, and contrasts between correct and incorrect sentences to enhance cognitive comparison. They also display contrasts between incorrect sentences that are frequently found in ESL compositions and those that reflect a possible influence of the participant's first language. If learners are able to associate such errors with the structure of their first language, such association may result in a cognitive scaffolding effect, whereby learners can use an already familiar linguistic basis to remember to avoid such grammatical errors in the future.

The ESL tutor is connected to a web-enabled database. Thus when users are engaged in a task, the time involved in reading a lesson or completing a task, the answers they choose or enter, and the feedback they receive are all recorded onto a database automatically. Using these capabilities for data collection, the program has been examined and tested to investigate the effectiveness of the program in terms of its design based on the corpus (Cowan, Choi, & Kim, 2003), its application of theory and design (Chapelle & Kim, 2003; Kim, 2003), its long-term effects (Lee, Choo, & Kim, 2003), and its feedback (Kim, 2005).

Participants

A total fifty five Korean ESL learners participated in two separate experiments in 2003. Twenty two advanced level English L2 learners (ten males and twelve females) participated in the first experiment in spring 2003. Thirty three intermediate and advanced level English L2 learners participated in the second experiment in fall 2003.

None of the participants in the first experiment participated in the second experiment and there was at least a six month interval between the two experiments. The participants were all graduate students from various fields of study at a mid-western university in the United States and were native speakers of Korean. All of the participants had taken at least one of the ESL writing courses offered by the university at the time of the study. Because they were from a single incoming group of students at the beginning of the school year, the participants in both experiments were assumed to be from the same population.

Procedures

The first experiment used a quasi-experimental design to test whether CALL instruction using negative evidence could help L2 learners eliminate certain L1 transfer errors such as overpassivization of ergative verbs (e.g., to *change*, to *increase*, to *sink*, to *happen*, to *occur* etc.), misuse of indefinite articles, and missing plural markers (Lee et al., 2003). Twenty two advanced level Korean ESL students who were enrolled in the ESL writing courses participated in the first experiment for seven weeks in spring 2003 and came to the language lab in small groups on different days to complete the experiment. All of them worked through three grammar topics such as passives, articles, and plurals in the CALL program after being given a two-page essay-type error correction task as a pretest, which was used as a posttest as well. The entire CALL instruction lasted about one hour. The participants returned one week and seven week after completing the CALL instruction to perform the same task with the passage to provide a long-term effect of the CALL instruction. Results from a preliminary analysis demonstrated that the CALL instruction had an impact on the participants' ability to identify and correct errors with both ergative verbs in passive voice and plurals, which indicates that the use of negative evidence in a CALL environment was effective in dealing with these problems as suggested.

Since the results from the first experiment with *the ESL tutor* were successful, a second quasi-experimental study was conducted to test the long term effectiveness of CALL instruction with more grammatical categories in fall 2003. Thirty three Korean-speaking ESL learners who were enrolled in one of the ESL writing service courses received the CALL instruction in the computer lab once a week for four weeks on four syntactic categories such as passives, articles, quantifiers, and demonstratives. They were randomly assigned to one of the two orders for taking the two CALL lessons each week. The participants were given a two page written passage for error correction as a pretest, which was also used as an immediate posttest one week after the final CALL instruction and a delayed posttest five months after the immediate posttest. The pre- and post-test result comparisons supported the hypothesis that the CALL instruction makes a significant improvement in Korean ESL learners' ability to detect and correct some of the persistent grammatical errors.

Data Collection

Learner data of the CALL instruction and the test results from the first experiment were used in the analysis for the first study on the effect of time spent on the tasks, gender, and learner proficiency on the learner performance. For the first study, the data of the participants who completed two grammar lessons (passives and nouns) were examined, but only the data from Section A to Section D were included in the statistical analyses due to the incompleteness of the activities in the other sections. Two females out of twenty two participants were excluded from the analysis because there were some missing time-tracking data for some sections

Only the data of the students who completed three grammar lessons on articles, demonstrative determiners, and passives in the second experiment were examined for the second study. If the records indicated that the participants had gone through any section more than once, only the initial attempts were considered for analysis to prevent memory effects on learner performance. After eliminating the participants who did not receive any corrective feedback, the researchers were left with twenty-one students' performance data out of thirty three.

ANALYSIS AND RESULTS

Study 1

To answer the first research question using a correlation and regression analysis, learner data gathered were recoded. The data include time spent on the lessons (unit: second), gender (10 males and 10 females using effect coding), the ESL writing course level (400 only, 401 only and 400 and 401 together using dummy coding) as a proficiency measure, and gain scores calculated from pretest and posttest scores. Since no other learner proficiency indicator such as a TOEFL score was available, the ESL writing course level was used as a learner's proficiency index in English writing. In summary, gain scores between pretest and posttest scores in the first experiment were used as a dependent variable and time spent on the lessons (time), gender, and ESL writing course experience were used as independent variables.

To see if there was any influence of time and gender on the improvement, an interaction model was employed to test the interaction between time and gender in addition to the main effect of time and gender. The correlation analysis shown in Table 1, indicated that time (M=30.61, SD=7.99, N=20) and the gain scores (M=5.60, SD=1.03) had a moderate positive relationship, $r(20)=0.403$, $p=0.078$, which is displayed in Figure 2. Interestingly, females seem to have spent more time on the lessons than males, $r(20)=-0.479$, $p=0.032$ as shown in Figure 3. However, as displayed in Table 2, multiple linear regression analyses found that there was no interaction between time and gender for the groups (t for $b_{\text{time*gender}}=-1.043$, $p=0.312$ at $\alpha=0.05$). To see if there was any gender effect controlling time on the improvement, another regression analysis model was applied, but no difference between two gender groups (t for $b_{\text{gender}}=0.245$, $p=0.810$ at $\alpha=.05$) was found.

Time effect in the coefficient table also shows no significance (t for b_{time} =1.715, p=0.105 at α=0.05).

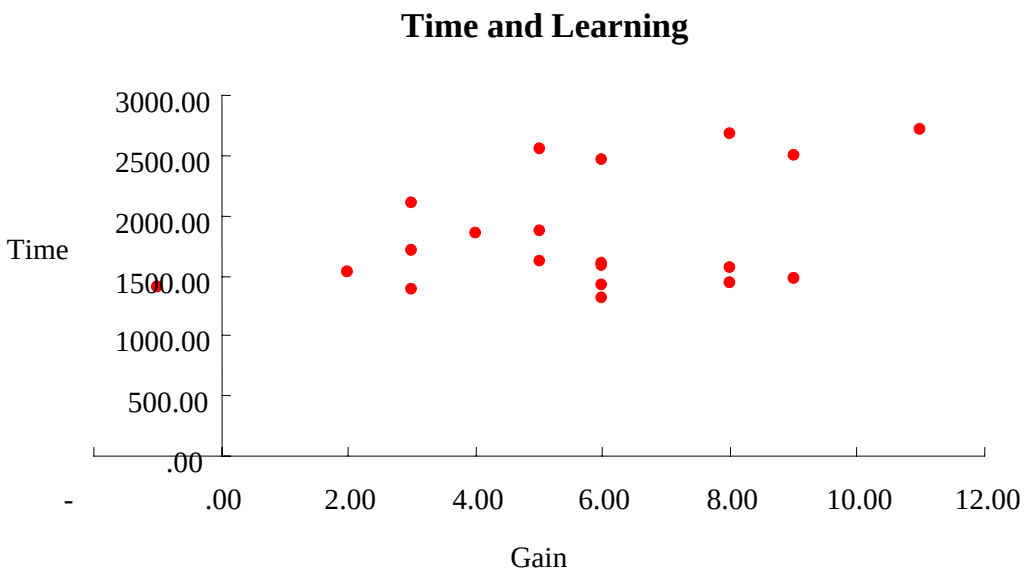


Figure 2. Time spent on the lessons and gain scores between pre and post tests

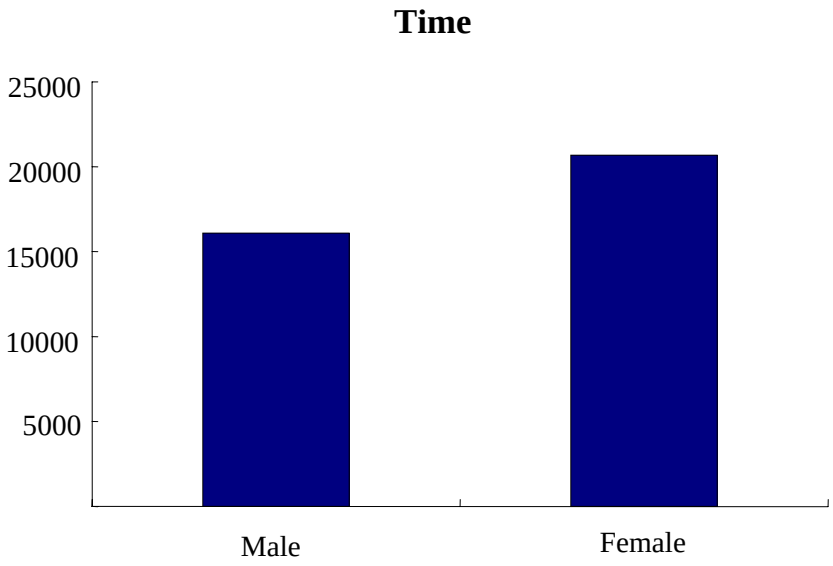


Figure 3. Time spent on the lessons per gender group

Table 1. Effects of time and gender on the improvement: Correlations and descriptive statistics (n=20)

Variable	1	2	3	4
1. Gain	--			
2. Time[a]	.40	--		
3. Gender[b]a	-.15	-.48*	--	
4. Time*Gender	-.21	-.52	.98	--
M	5.60	30.62a	.00	-3.74
SD	2.82	8.00	1.03	32.19

[a]Time: unit = minute, [b]Gender: 1 = male, -1=female

* $p < .05$

Table 2. Summary of multiple regression analysis for time and gender predicting the improvement (n=20)

Variable	Model 1			Model 2			Model 3		
	B	*SE B*	β	*B*	*SE B*	β	*B*	*SE B*	β
Time	.14	.08	.40	.15	.09	.43	.13	.09	.36
Gender				.17	.69	.06	3.02	2.82	1.10
Time*Gender							.00	.092	-1.10
R^2	.	.16			.17			.22	
F for change in R^2		.08			.81			.31	

The marginal correlation between time and gain scores suggests that there might be a curvilinear relationship between these two variables. Therefore, the time variable was recoded into three variables (time1, time2, and time3) to see if there was any quadratic or cubic relationship between time and gain scores. In order to do a polynomial regression analysis with quadratic and cubic terms of time variable, a criterion of very low Tolerance was chosen so that the quadratic and cubic terms could be entered into the equation in SPSS. In the second regression model, the proportion of variance accounted

for by the linear, quadratic, and cubic regression were only 16%, 8%, and 7% each (Significance of F change test for linear=0.078, for quadratic=0.212, and for cubic=0.249). Hence no significant linear, quadratic, or cubic relationship for time on the improvement was found at α=0.05 level.

To examine the relationship between ESL writing course taking experience based upon proficiency in writing and the improvement, a regression analysis was conducted, but there was no difference in group means between the three different groupings of ESL courses (400 only, 401 only and 400 and 401 together) which were used to indicate educational experience (Sig. F change=0.608 at α=0.05). This indicates that enrollment in different ESL writing courses did not seem to affect the learners' success in the use of a CALL program.

Study 2

The purpose of our second study was to investigate the effectiveness of three types of correctional feedback built into an error correction task in *the ESL Tutor* on learner performance by examining previously collected data. Since Section D simulated an interactive editing session with a tutor, students had multiple opportunities to identify and correct errors within a passage. The program provided multiple feedback statements for both finding and correcting errors. Since most students had successfully found errors, our focus was on the types of feedback students received for each attempt to correct an error. If the student's input matched the answer recorded in the database, then he/she received a positive feedback comment. If the student did not enter an appropriate correction, a correctional feedback appeared on the screen.

There were three types of correctional feedback provided in *the ESL Tutor*. Firstly, the "expected" type of feedback statement appeared if the student's 'erroneous' correction matched a predicted set of errors—also known as "prepackaged feedback" (Brandl, 1995, p. 208)—based on the corpus analysis of common errors found in essays from ESL writing courses. This type of feedback also was designed to encourage noticing and focus-on-form. Secondly, a "try again" type of feedback statement was displayed if the student's response did not match any predicted responses during the first try. Such type of feedback let the users know the answer was wrong and provided them with preset instruction such as checking spelling and reminding them of the lessons in previous sections. Thirdly, a "generic" type of feedback statement appeared during the student's second attempt to correct the error when his or her correction did not match any of the predicted errors. It provided a repetition of the erroneous attempt in addition to a right/wrong response. Thus, the participants had two attempts for each error that was correctly highlighted. Specific examples of the three types of feedback are shown below:

1. Try Again Feedback

Prompt: Although we have had a lot of success with this program, it is hard to know how long it will be lasted.

Student highlights:	be lasted
Portion to correct:	will be lasted
Attempt 1:	will lasted
Feedback:	*Sorry, your answer is wrong. Check your spelling, and make sure your correction follows what you've learned in the previous sections. Try again.*
Attempt 2:	will last

2. Expected Feedback & Generic Feedback

Prompt:	It is ridiculous that most women in developing countries are suffered from poverty.
Student highlights:	are suffered
Portion to correct:	are suffered
Attempt 1:	suffered
Feedback:	*suffer*ED*?* (Expected Feedback)
Attempt 2:	are suffer
Feedback:	*That's not right, either, Jungsoo. Let's try once more. Hint: Can you say ARE SUFFERED?* (Generic Feedback)
Attempt 3:	suffer

The users' responses and the feedback they received were stored in the database, so that the instructor could monitor their progress and increase the quality of feedback. The data analyzed in this second study were the type of feedback students had received (expected, generic, and try again), and the score that they had obtained. A descriptive account of the data is provided in Table 3 in terms of the mean time spent on each task and the mean score for each grammatical category. The difficulty of each grammatical category is evident from these data as is the low mean scores for the article category in both sections B and D. Article errors are known to be a difficult area for Korean students.

Table 3. Descriptive analysis of the first four sections of the participants' performance data

Grammatical Category	Section A (seconds)	Section B (%)	Section C (seconds)	Section D (%)
Article	163.24	53.57	131.67	86.05
Demonstrative Determiners	19.48	94.29	34.81	91.29
Passive	60.14	77.14	302.90	91.48

Table 4. Performance data for section D in terms of feedback

Type	Expected		Try Again		Generic		Total			
Category	Yes	No	Yes	No	Yes	No	Yes	No	Sum	Proportion
A	14	4	18	14	10	3	42	21	63	66.67%
D	19	7	20	20	5	9	44	36	80	55.00%
P	25	7	14	7	9	1	48	15	63	76.19%
Sum	58	18	52	41	24	13	134	72	206	65.05%
Proportion	76.32%		55.91%		64.86%					

Note. A = Article, D = Demonstrative Determiner, P = Passive

Table 4 shows a descriptive account of the performance information from Section D in terms of feedback. Descriptively speaking, 21 students received 206 feedback statements. Of those, students were able to provide a correct answer after 134 feedback statements (65.05%). Most correct answers in terms of grammatical categories were given in the passive category (76.19%). In terms of types of feedback, the expected type of feedback appeared to generate the most correct answers (76.32%). The generic type of feedback seemed to be the next feedback type that led to more correct answers (64.86%), and the try again type of feedback seemed to produce the least amount of appropriate corrections (55.91%). In conclusion, this simple descriptive analysis indicates that out of all the correctional feedback displayed during the experiment, the "expected" type of feedback resulted in the most correct answers for the learners' who did not supply the proper correction.

The response patterns varied because some participants were able to correct the errors in their first try while others did not until they received one or two corrective feedback statements. Since we wanted to examine the effectiveness of the corrective feedback types, we naturally focused only on the attempts where students had failed to provide a correct answer at least in the initial trial. Because the type of feedback was related to the number of trials that were non-independent events, we used a more advanced categorical data analysis called *survival analysis method* in order to see which type of feedback led the students to enter more correct answers. This method, although generally used in the medical sciences, can analyze events associated with time. In this study, the event of interest is a correct response following a corrective feedback, and the event takes places within two trials. Furthermore, we assume that the event is affected by the type of feedback participants receive. This method was suggested because it considers censored data, that is, data that is unobservable. In this particular CALL context, corrective feedback was only observable within two trials where students were incorrect in their first try, and most students were able to supply the correct answer more often than not.

A univariate analysis was conducted for each of the categorical predictors to see if the predictor was relevant to the model. The log-rank test of equality across the strata "feedback" returned a p-value that is smaller than 0.0001, and was therefore included in the model. The graph in Figure 4 shows the *try again* feedback on the left ending at time=1, the *expected* type of feedback in the middle which stretches out across time=1 and time=2, and the *generic* type of feedback which ends at time=2. In fact, Figure 4 accurately represents where each feedback was presented given the limited trials. The *expected* type of feedback statement is designed to appear at any one of the two attempts, whereas the *try again* feedback was designed to appear only on the first attempt, and the *generic* feedback statement was designated only on the second attempt in the absence of a predicted correction.

Another univariate test was conducted with the score strata to see whether it could fit the model. Again, the p-value for the log-rank test of equality of the strata "score" resulted in a p-value smaller than 0.0001. Thus, the strata *score* was included in the model.

A further test was conducted to the feedback variable because it included three levels. Each of the levels using proc phreg was examined by including dummy variables. The result is that the feedback variable was significant at $p<0.0001$. (Note: To use proc phreg, the text variables for feedback e, t, g were replaced by numerical variables 1, 2, and 3) So, the model was created using *feedback* and *score* as main effects. Since the focus of this analysis was to see how different feedback types can influence the scores, we decided to test a possible interaction between the feedback and score variable. The interaction variable *feedscore* turns out to be non-significant, and thus will not be included in the model.

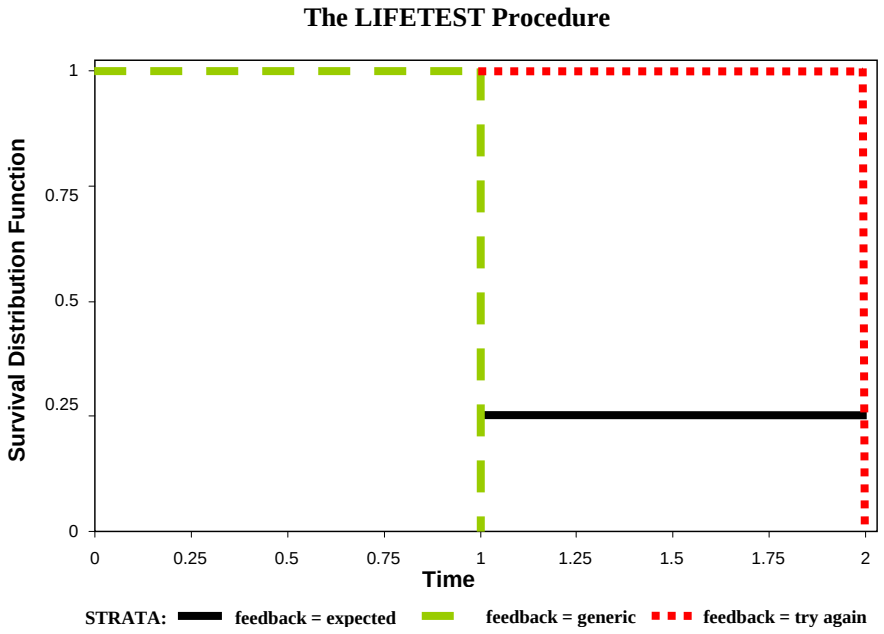


Figure 4. The LIFETEST procedure

Table 5. SAS output: Analysis of maximum likelihood estimates

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
feedback	1	-0.228	0.08895	6.5697	0.0104	0.796
Score	1	-0.00486	0.14534	0.0011	0.9733	0.995

The hazard ratio for feedback as illustrated in Table 5 indicates that as the feedback type moved from the *expected* type to *try again* type, and from the *try again* type to the *generic* type, the rate of providing a correct answer by the participant decreased by 20.4% (=100% - 79.6%). The score variable high p-value of 0.9733 did not seem to warrant an interpretation. Thus, it can be concluded that the *expected* type of feedback resulted in the most correct answers, which seems to corroborate the hypothesis that students respond differently depending on the type of feedback, and that response-specific feedback is likely to be worth the effort it takes to develop it.

DISCUSSION

Based upon the results from the regression analysis, time spent on the program, regardless of the users' gender, did not influence the learning. However, a marginally positive correlation between time and learning improvement was found. This suggests that though time on task in the CALL software was likely to have a positive relationship with learner improvement, it may not be a critical predictor variable for successful learning as use of the CALL software, which is in line with the findings of Hegelheimer & Tower's (2004) study. In addition, female users in the current study seemed to be more patient (or slower) to try the program than male users with no different learning effect, which accords with Astleitner & Steinberg's (2005) conclusions that gender may only play a very minor role in web-based learning. Finally the previous or current ESL writing course experience based upon the learner proficiency level did not seem to play a role to affect the learning via the CALL program in this analysis. In other words, the program itself may be effective enough to be used across a range of levels of ESL learners.

These findings must be interpreted in view of the possibility that other variables should perhaps be controlled for more accurate analysis. For the current study, no other personal information such as length of stay in the U.S., majors and grades from the ESL course and user perception about the program was collected. Based on participants' comments

we believe that some students might not like some of the program features, so they simply did not pay attention to some of the sections with the features they did not like. Two female users skipped some of the sections for unknown reasons. If some of these variables had been included with larger number of observations, different results might have been found. Although not many interesting results were found from the multiple regression analyses in general, the results still give us some insight into the learner performance with CALL instruction and variable selection for future experiments.

In terms of the effectiveness of the types of feedback used in this study, the *expected* type of feedback was found to lead to the highest rate of resulting correct responses. Such feedback messages simulate the type of written responses teachers provide in student compositions. Because such errors were identified based on the analysis of learner writing, it was possible to preemptively enter a list of possible student responses for a given item. Although this approach is criticized by developers of natural language processing (NLP)-based CALL programs (e.g., Nagata, 2002) due to the laborious nature of entering such feedback and its restrictive applicability, it seemed to be useful and effective for targeting certain persistent errors.

The survival analysis method used in Study 2 seemed to support the results in the descriptive study of the feedback types from Kim's (2005) previous study, where the expected type of feedback resulted in the most correct responses. Analyzing the effectiveness of various types of correctional feedback on learning can be challenging because feedback will only appear when the student produces an error. This is especially true of the performance of advanced learners who produce more correct responses than errors.

CONCLUSION

In this paper we have shown how learner performance data from a CALL program can be statistically analyzed, which illustrates that learner data can reveal important information about learner performance and the effectiveness of the CALL program in terms of its tasks and feedback. Although the most important goal in designing a CALL program should be to enhance learning, designers should consider ways to embed tracking features that will also enhance research in SLA. Developing ways to count the frequency of a learner's use of various CALL features may help determine the effectiveness of the interface design. However, even within a single task, various cognitive indicators might be measured. When users change the answers on a single item, this may indicate a learner's weak confidence in the content presented. Such wavering behavior can be easily detected and accounted for in a learner profile. Although time spent on task or on the entire program may not be a good predictor of success in learning, the time that each feedback is displayed may be measured to see if the learner actually reads the feedback. Buttons that ask about learner confidence about an item may be used as well. For example, learners may choose an answer, and instead of clicking a confirm button to submit the answer, multiple buttons may be presented so that learners can go to the next

item by clicking on a button that indicates how confident they are in their choices. Such information can be used pedagogically to provide lessons or tasks that encourage individualized learning.

REFERENCES

Astleitner, H., & Steinberg, R. (2005). Are there gender differences in web-based learning? An integrated model and related effect sizes. *AACE Journal, 13*(1), 47-63.

Beaudoin, M. (2004). Educational use of databases in CALL. *Computer Assisted Language Learning, 17*(5), 497-516.

Brandl, K. K. (1995). Strong and weak students' preferences for error feedback options and responses. *The Modern Language Journal, 79*, 194-211.

Chapelle, C. A. (2005). Interactionist SLA theory in CALL research. In J. Egbert & G. M. Petrie (Eds.), *CALL research perspectives* (pp. 53-64). Mahwah, NJ: Lawrence Erlbaum Associates.

Chapelle, C. A. (2003). *English language learning and technology*. Amsterdam: John Benjamins.

Chapelle, C. A., & Kim, D. (2003, October). Linking SLA theory to technology based tasks. Paper presented at the Conference on Technology for Second Language Learning, Ames, IA.

Colley, A. & Comber, C. (2003). Age and gender differences in computer use and attitudes among secondary school students: what has changed?, *Educational Research, 45*(2), 155-165.

Cowan, R., Choi, H. E., & Kim, D. H. (2003). Four questions for error diagnosis and correction in CALL. *CALICO Journal, 20*, 451-463.

Glendinning, E., & Howard, R. (2003). Lotus ScreenCam as an aid to investigating student writing. *Computer Assisted Language Learning, 16*, 31-46.

Hegelheimer, V., & Tower, D. (2004). Using CALL in the classroom: Analyzing student interactions in an authentic classroom. *System, 32*, 185-205.

Kim, D. (2003, March). Designing and evaluating in CASLA. In C. Chapelle (Discussant), *Approaches to computer assisted second language acquisition (CASLA) research.* Symposium conducted at the meeting of the American Association of Applied Linguistics, Arlington, VA.

Kim, D.-H. (2005). *Finding feedback: An effective computer-assisted language learning model for second language acquisition.* Unpublished master's thesis, University of Illinois at Urbana-Champaign, Urbana, IL.

Lee, G., Choo, J., & Kim, D.-H. (2005, April). *Investigating the effectiveness of a CALL program designed to improve editing*. Paper presented at the SLATE Symposium, Urbana, IL.

Mitra, A., Lenzmeier, S., Steffensmeier, T., Avon, R., Qu, N. and Hazen, M. (2001). Gender and computer use in an academic institution: report from a longitudinal study, *Journal of Educational Computing Research, 23*, 67-84.

Nagata, N. (2002). BANZAI: An application of natural language processing to webbased language learning. *CALICO Journal, 19*, 583-599.

Skehan, P. (2003). Focus on form, tasks, and technology. *Computer Assisted Language Learning, 16*(5), 391-411.

Van der Linden, E. (1993). Does feedback enhance computer-assisted language learning? *Computers and Education, 21*(1), 61-65.

Wible, D., Kuo, C.-H., Chien, F.-y., Liu, A., & Tsao, N.-L. (2001). A Web-based EFL writing environment: integrating information for learners, teachers, and researchers. *Computers in Human Behavior, 37*, 279-315.

Schulze, M. (2008). Modeling SLA processes using NLP. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 149- 166). Ames, IA: Iowa State University.

Modeling SLA Processes using NLP

Mathias Schulze
Department of Germanic and Slavic Studies / Centre of Arts and Technology
University of Waterloo, Ontario, Canada

Attempting to understand and to capture the complex and dynamic nature of language learning processes is a non-trivial task for researchers in Second Language Acquisition (SLA) and Computer-Assisted Language Learning (CALL). After sketching major developments in SLA and student modeling for Intelligent CALL—the intersection of Artficial Intelligence and CALL—this paper proposes a conceptual framework of the dynamic language proficiency of students using the example of the Mocha project. The main project goals are outlined. Dynamic Systems Theory and Construction Grammar are motivated as the theoretical foundation of our thinking about SLA and student modeling.

Individualization has been praised as one of the major strengths and advantages of computer-assisted language learning (CALL). Advocates of CALL frequently mention the value of students working at their own pace and receiving feedback immediately and—sometimes—based on an individual context. However, individualization from this perspective often means having the student work individually using a tutorial CALL package or simply lack of instructor control. Individualization in this sense does not imply that individual characteristics of the students are considered and that the computational learning environment is tailored accordingly. Often it does not even mean that the prior learning path—learning events with their contents and the students' achievements—is recorded and has an influence on such decisions as which learning objects are presented or what kind of feedback is provided. I will argue in the course of this paper that a non-mechanical and more humanistic approach to individualization in CALL can only be achieved when the CALL system comprises a student model, a model which is informed through the analysis of learner texts.

Let us begin with a cursory look at trends in second language acquisition (SLA) research from which one might seek guidance for development of student models and note the assessment problem that arises in this context. The subsequent discussion of student models will show how different modeling techniques correspond to different approaches in SLA and will highlight their strengths and limitations that pertain to individualization. The main part of this paper will sketch a new approach to modeling language learning, an approach which relies on aspects of Dynamic Systems Theory (DST).[i] The consequences such a modeling approach has for the underlying natural language processing (NLP)

techniques will be illustrated through the description of a project currently underway which aims to conceptualize and construct a student model for early and intermediate learners of German at a Canadian university.

APPROACHES TO SECOND LANGUAGE ACQUISITION

Our current understanding of second language acquisition processes can be said to have started with Lado (1957) and his 'contrastive hypothesis.' The stronger version of this hypothesis claims that areas of learning difficulty can be predicted on the basis of typological analyses of the language acquired first (L1) and the language learnt or acquired later (L2). It then goes on to claim, for example, that if structural features of L2 differ greatly from comparable features in L1, learners of L2 will make errors when using linguistic structures with these features. This hypothesis had—and still has—some intuitive appeal. However, its predictive power is limited: learners make errors and have some difficulty with language phenomena that are very similar in their L1 and L2 (e.g., the learning and un-learning of cognates and false friends) and, vice versa, few or hardly any transfer-induced errors occur in some areas which are different in L1 and L2 (e.g., German learners of English usually have little trouble spelling English words in spite of the fact that the orthographic systems of the two languages are radically different). This shows that typological differences between L1 and L2 certainly influence the nature of the language learning process, but this variable is not an exclusive and sufficient predictive factor of language learning behavior and success—not even a dominant one—because the contrastive hypothesis in its reliance on comparing two language systems—two langues—fails to consider the language learner.

SLA researchers made an attempt to address these limitations by focusing on the learner language through error analysis (Corder, 1974, 1981). This approach did indeed improve our understanding of language learning processes by identifying areas of genuine difficulty for certain groups of learners. However, the exclusive focus on errors—the negative results of an individual learner's efforts—led to the exclusion of error avoidance strategies (Schachter, 1974) learners employ because these could not be detected on the surface in the analysis of an individual text. The focus on errors also meant that learner language was described as flawed, shortcomings were emphasized, and learner success and interesting, creative, and meaningful aspects of learner utterances and texts were ignored. Thus, the learner's identity construction in text—their image as depicted in the what and how of the text—was reduced to the classificatory description of selected negative features. Social and individual characteristics of the learner as text producer did not feature in error analysis.

The sole concentration on learner errors was overcome in interlanguage analysis (Selinker, 1974, 1992). All linguistic structures in texts by individual learners over time were considered. This interlanguage continuum—both across groups of learners and for each learner over time—was conceptualized as a variety space (Klein, 1986; Klein & Dittmar, 1979; Klein & Perdue, 1992). Each variety was assumed to have its own system

of rules. Thus interlanguage was and still is described as systematic, a conceptualization which provides a basis for a computational implementation of the interlanguage grammar. These language systems consist of rules which are identical to rules in L1 or other previously acquired languages (language transfer) or in L2 (acquired rules) as well as rules which are specific to this particular variety (e.g., through overgeneralization and simplification). The often purely structural focus of interlanguage studies has been extended more recently to include the pragmatics of interlanguage (e.g., Kasper & Rose, 2002). Despite the theorized systematicity of individual interlanguages, considerable variability is also observed, and as a consequence, modeling interlanguages has proven more difficult than one might initially expect.

Perhaps even more problematic for conceptualizing student models in CALL, interlanguage studies have largely ignored individual learner differences (Dörnyei, 2005) and concentrated on groups of learners. Even when individual differences are studied in SLA research, researchers often looked at one variable after attempting to statistically or experimentally eliminate all others. This contributed to—what was perceived by some as—a dominance of quantitative approaches in SLA (Firth & Wagner, 1997). The concentration on individual variables yielded interesting insights into important aspects of language learning processes and language learners and instructors. However, the exclusion of other variables—often minor ones—resulted in these (quantitative) studies presenting contradictory results (Larsen-Freeman, 1997).

It is likely that contextual differences in these studies have a disproportionate effect on their findings. In other words, quantitative studies which should be generalizable, may be impossible to replicate because of seemingly innocuous contextual variables. In contrast, qualitative (case) studies made every attempt to consider all variables which might possibly influence the learner, the learning process, and the learning outcomes as well as the instructor. Such studies posed challenging new questions and highlighted interesting factors which influence language learning, but they are difficult to generalize and even more difficult to apply and implement in a computational context such as student modeling. Comprehensive, generalizable, and robust findings in SLA can provide a solid basis for student models in CALL.

Addressing this dilemma of quantitative vs. qualitative research methods, integrative approaches to SLA have been suggested more recently (de Bot, Lowie, & Verspoor, 2005, 2007; Ellis & Larsen-Freeman, 2006; Larsen-Freeman, 1997, 2000, 2003). These approaches, which can be referred to under the umbrella of Dynamic Systems Theory (DST) have in common that they view language learning as a dynamic system: a system which changes over time and is changed by individual speakers in communicative events. The DST perspective conceptualizes second language acquisition as:

- nonlinear, i.e., it incorporates spurts, progression, plateaus, fossilization, and retrogression, …;
- nonperiodic, i.e., segments of the language learning process will not be repeated

and will not recur, but smaller segments of the learning curve might be similar to larger segments—a self-similarity of the plot of learning events which has often been described as fractality;

- nonmonotonic, i.e., the speed of L2 acquisition and / or attrition varies over time;
- complex, i.e., a large number of variables will have to be considered at any given point in time, their correlations have to be taken into account as well as the changes they undergo through the interaction with other variables and over time.

From a DST perspective, initial conditions—and even tiny differences in these initial conditions—often result in large differences in the end state of the system due to the complex nature of the dynamic system. Based on observation of this complex system, it appears to be impossible to predict the quality of the end state of the system, e.g., it is impossible to predict after observing an early language learner whether she will ever reach near-native proficiency and what this proficiency will look like for her. However, given meaningful data points over time, it should be possible to predict the nature of the next state as a (mathematical) function of the path through prior states. To use an example from another dynamic system, it is possible today to make relatively reliable predictions about the weather of the next twelve to twenty-four hours, based on weather data of the last hundred years or so, but it is impossible to predict the moment when the weather system is coming to a rest (when the weather will not change any longer) and the quality of that end state nor is it possible to predict the exact nature of the weather at some point in time in the more remote future (Lorenz, 1993). Similar claims can be made about language learning: It is our hypothesis that it is possible to make valid predictions of the next state of a particular learner's language learning system, but it will be impossible to make predictions about the state of the language learning process in the more remote future.

The DST approach to SLA is relatively new with a growing number of researchers entering into the discussion. There are promising results however in other social sciences: cognitive psychology and first language acquisition (Hollich, 2000; Hollich et al., 2000; van Geert, 1994, 1997, 1999, 2000; van Geert, Verhoeven, & van Balkom, 2004), pedagogy (Haggis, 2005), and bilingualism (Herdina & Jessner, 2002). The application of the underlying philosophical approach in DST to SLA—the consideration of a multitude of variables[ii] (or at least their concatenation) in context—promises a framework for a more integrative conceptualization of language learning. Moreover, DST has a mathematical basis, which may provide a basis for its computational implementation and therefore an impetus for student modeling in CALL.

BRIEF EXCURSION INTO LANGUAGE ASSESSMENT

A DST conceptualization of SLA may provide a promising basis for student models, but on the surface it presents a challenge for testing procedures. If we want to measure complex variables such as proficiency or even communicative competence then we need

to gain an understanding of the complexity of the learner and of the language they use. Classical testing theory relies on calibrating the difficulty of decontextualized (often sentential) test items and to measure students' ability as a reverse proportional function of the difficulty of the most difficult items correctly answered. However, resulting scores provide only a holistic estimate of proficiency or only one aspect of proficiency (see Jang, this volume). In order to develop assessments that can provide more fine-tuned diagnostic information that can play a role in individualized CALL, we need to be able to model learners in their complexity. For example, if the student model can predict the nature of the next language learning system state the learner is likely to be in, we can present the learner with a test item which corresponds to that state.

Given our notion of language learning as a dynamic system, we are able to view a language test as a complex language learning event which 'interacts' with prior language learning events and whose test items interact with each other. A detailed linguistic analysis—in computer-assisted language testing preferably a computational analysis—is a necessary prerequisite for the analysis of test results as a whole. And again, modeling the language learning of individual students in context is a key to successful adaptive (diagnostic) testing in an integrative, communicative approach to SLA. What could such models look like?

APPROACHES TO STUDENT MODELING

Student models gather and structure information about the student's knowledge. In computational terms, a student model can be defined as a data structure that contains information about the student. "But we cannot directly observe what a student does and does not *know*; this we must infer, imperfectly, from what a student does and does not *do*" (Mislevy & Gitomer, 1996, p. 253). Thus, a distinction has to be made between a surface level student model which "represents the scheduled problem solving plans and applied procedural knowledge" and a deeper level student model which "must infer and model the student's belief by interpreting the surface level student model" (Villano, 1992, p. 469).

The simplest way of maintaining such a data structure is by recording performance data in the form of scores. These scores, however, do not contain any information about the kind of knowledge that has been acquired; they only reflect how much knowledge has been gained (Gisolfi, Dattolo, & Balzano, 1992, p. 329). For example, looking at the score of a simple grammar test conducted in the foreign language classroom, we could note that student X answered 80% of the questions correctly. This result might be about 20% higher than that of a comparable previous test (probable knowledge gain). On the other hand, if we looked at this score later, we would be unable to ascertain what exactly the knowledge items were the student had acquired or not. Most student models capture the knowledge state(s) of the student relative to the domain of learning. Far fewer incorporate individual characteristics of the learner (Milne, Shiu, & Cook, 1996) because individual characteristics are much more difficult to obtain (Mabbott & Bull, 2004).

Student models can be used in a number of ways (Elsom-Cook, 1993): corrective (providing tailored feedback), elaborative (extending the knowledge of the student), strategic (guiding decisions on teaching interventions), diagnostic (to determine the knowledge state of the student), predictive (anticipate future student behavior) and evaluative (assessing the level of student achievement). The diagnostic and evaluative uses can be incorporated into CALL programs or can function in language assessments.

A number of different approaches of creating and maintaining a student model facilitate the gathering of detailed information about the language learner and her learning process and avoid the exclusive reliance on score-keeping. Here, I will briefly discuss the bug library technique, the model tracing technique, and constraint-based student modeling.[iii]

The Bug Library Technique

The bug library technique comprises error descriptions of student errors and their explanations. Murray (1999, p. 99) distinguishes two different ways of recording student bugs: runnable models (student knowledge as subset of expert system rules plus some buggy rules) and overlay models (assign competency or probability to different rules according to inferences the system has made). A version of the bug catalogue is the so-called perturbation approach. Reyes (1998, p. 330) suggests this approach to student modeling for the domain of Pascal programming. This technique does not rely on a set of buggy rules, that is, anticipated student errors. Instead, it uses transformations of rules that the expert system possesses. She applies perturbation, that is, a set of meta-rules which modify operators, delete sub-expressions, exchange operands and alter variables (Reyes, 1998, p. 330). Perturbation or other modeling techniques discussed later are applied because the error descriptions are very expensive to create since they are built on empirical analyses of errors previously encountered. The error libraries are also restricted in that they are not transferable from one student population to another. Errors often occur because the student applied a similar rule, schema, or pattern to a new problem,[iv] or because the learner employed an existing correct rule which is not appropriate for the problem or the context of the problem at hand (Burton & Brown, 1982). How problematic this is in language learning and teaching becomes apparent if one considers that the number of utterances in any language is infinite and because each of these utterances could at least contain one error, the number of erroneous utterances is also infinite. Programs with a built-in expert system—in Intelligent CALL (ICALL) systems, with their application of artificial intelligence techniques to CALL, this is usually the case with an in-built natural language parser—have been based all too often on the assumption that the student's knowledge is simply a subset of the knowledge of the expert system (e.g., Cerri, Cheli, & McIntyre, 1992). Accordingly, the main function of the system is to impart the complementary subset of facts and rules onto the student. This is, of course, a fallacy and simplifies the teaching and learning process (Burton & Brown, 1982, p. 51). If one compares facets of the bug library approach, it is apparent that there are interesting overlaps with the contrastive and error analyses approaches in SLA. Thus, it is difficult to apply this student modeling technique in the context of current discourses in SLA.

The Model Tracing Technique

The model tracing technique monitors each step the student takes in the problem solving process instead of attempting to infer from final answers. "The student is modeled as the set of rules which matched his or her steps in the traced performance" (Ohlsson, 1992, p. 433). The technique thus depends on a set of correct and incorrect rules and has to rely on anticipating the rules that might get violated. Tasso, Fum and Giangrandi (1992) developed a version of this technique, that is, of backward model tracing. It was implemented for a later version of ET (English Tutor), which concentrates on verb conjugation in English. Backward model tracing utilizes all techniques of model tracing, but does not rely "on an a priori established catalogue of correct and incorrect productions but is able to dynamically generate mal-rules necessary to explain the student performance" (Tasso et al., 1992, p. 154). The student input is compared with a version of the same utterance which is generated by not just relying on the expert system, but also on the information already contained in the student model. Accordingly, the actual student performance is compared to the expected student behavior as predicted by the learner model. Tasso et al. (1992) state:

> If the two answers are equal, the Modeler assumes that the student has applied the same knowledge utilized in the simulation process and this constitutes a useful piece of information for discriminating among possible hypotheses still active from preceding cycles. On the other hand, if the two answers are different, the Modeler executes the two analyses of commission [application of an inappropriate rule] and omission rules [ignoring of a necessary rule], which will eventually produce new hypotheses about the student knowledge. (Tasso et al., 1992, p. 158)

It is not surprising that such an application of model tracing was to language learning in Krashen's (1982) narrow sense of learning (vs. acquisition), e.g., to the learning of individual grammatical rules. Only such learning procedures can be described algorithmically, whereas the production of chunks of discourses in meaningful communication cannot be broken down into a neat sequence of steps the learner or text producer has to follow. In other words, if the system is intended to model the different steps students take when learning grammatical knowledge items, model tracing appears to be a viable option, whereas if the CALL system is intended as a tool or tutor (Levy 1997) in communicative language teaching, it is not necessarily possible to establish and trace an ordered sequence of deterministic steps the learner took or will have to take. This limitation led ICALL researchers to the exploration of constraint-based approaches, which came from formal linguistics and robust parsing, in order to decrease the reliance on (error) anticipation.

Constraint-Based Modeling

Constraint-based modeling was originally proposed by Ohlsson (see e.g. Mitrovic, 1998, p. 415). This approach concentrates on learner errors and attempts to correct them. It presupposes that diagnostic information is not attained from the (intractable) learning problem solving process the student undergoes. Instead it is obtained from the problem state at which the student arrives or the final results. The constraint-based approach has

similarities with the bug library technique described above. Both start with a knowledge base in the expert system. The bug library technique generates possible deviant solutions with meta-rules that transform rules and facts from the expert system. The constraint-based approach relaxes these constraints during the analysis in order to determine the rule(s) that might have been violated. Each constraint consists of a relevance condition that, in turn, determines when to apply the satisfaction condition (Ohlsson, 1992, p. 437). For example, if the parser finds what could be a direct object, then the phrase in question needs to be marked for accusative case. If this constraint is relaxed, however, the parser would successfully parse the sentence containing the direct object even though the object might be marked with the dative case. The parser would record that a phrase that should have been accusative-marked was actually dative-marked to later on have, for example, a basis to provide feedback to the learner about the deviant case-marking. The constraint-based modeling approach is very efficient because it does not rely on the anticipation of student errors as an intractable problem (Heift, 1998; Heift & Nicholson, 2001; Heinecke, Kunze, Menzel, & Schröder, 1998; Menzel, 1988, 1992a, 1992b; Menzel & Schröder, 1998; Schröder, 2002; Schulze, 1998, 1999, 2001; Vandeventer, 2001). However, it poses a set of different problems because it has to work with an immensely large search space. For example, parse forests, that is, all syntactic trees representing one and the same sentence or text fragment, get increasingly larger depending on the number and kind of constraints that can be violated. This potentially prolongs the analysis of student input. But, more importantly, it requires a selection of the most appropriate analysis of this input for feedback generation and for the maintenance of the student model. This, in itself, is a very complicated and time-consuming task.

This modeling approach bears some similarity with earlier conceptualizations of interlanguage (Selinker, 1974, 1992) in that it starts off with the assumption that the ultimate goal of language learning—the target variety—is the standard variety used by L2 native speakers and that surface structures of this variety are the main ‘content’ for acquisition. This modeling approach pays close attention to interlanguage processes such as transfer, overgeneralization, and simplification, but would have significant problems with what Corder described as errors of appropriateness (Corder, 1974), i.e., how communicative intentions were met successfully or otherwise.

We are not aware of any other modeling approaches that have been applied to CALL successfully (Heift & Schulze, 2007). We could assume that it would be necessary to rely on machine learning approaches to capture the dynamics and the complexity of the second language acquisition processes, but this remains a question which will have to be answered by future research although some problems of modeling learning in such domains have already been addressed in student modeling research.

General Problems with Student Modeling

McCalla et al. (2000) believe that learner modeling "should be easier than in the past given the vast amount of information that will be available about learner interaction in the emerging information technology intensive world" (p. 61). Accordingly, this

conceptualization of a student model views modeling as a computation, that is, a continuous, fragmented process rather than a data structure. As a result, the immense development costs of learner models, which others have tried to achieve with generic student models, are reduced. The challenge of this approach to student modeling, however, does not primarily lie in the information collected about the learner, but in the fact that the large amount of information available needs to result in a coherent student model in spite of its many facets and traits. Student models are more complex than other user models because misconceptions and inconsistencies in the student's knowledge have to be considered. Mitrovic et al. (1996) identify four different sources of this noise: Inconsistent student behavior, dynamic and nonmonotonic nature of human learning, ambiguity of possible explanations for the observed behavior and indeterminacy of student answers. Certain aspects of observable student behavior can only be described as stochastic. As discussed in the previous section, we are addressing this challenge of complexity and nonlinearity through our reliance on DST approaches.

The following are a few possible reasons for students' inconsistent beliefs:

- In language learning, students are often testing hypotheses (Output Hypothesis - Swain, 1985) and they expect feedback on these attempts. There are instances when students deliberately employ a wide variety of surface structures which—according to their hypothesis—convey the same meaning to learn which of these will be accepted.

- Students are often unable to consider all relevant issues simultaneously when solving a problem because nobody has full access to one's full body of knowledge at all times. Considering issues such as conflicting communicative (sub)goals, meaning and form of lexical items, word order rules, subject-verb agreement, subcategorization of verb arguments, case-marking of noun phrases, … all at the same time results in nonlinear, inconsistent behavior.

- Students react adversely to external pressures. Such pressures often stem from testing situations but also factors such as fatigue, lack of motivation can play a role. It is also possible that one source of inconsistent student behavior can be attributed to the fact that students are using computers for language learning. Low levels of computer literacy and lack of typing skills can adversely affect language learning behavior. On the other hand, a preference for working with computers, a positive attitude to computer-mediated communication, for example, can also be conducive to language learning.

These are some of the factors which will have to be considered when conceptualizing a student model for language learning.

Conceptualization of the Mocha Student Model

We are basing our approach to student modeling in the Mocha project[v] on an integrative and balanced approach to SLA (Larsen-Freeman, 2000) by relying heavily on DST

(Lorenz, 1993; van Geert, 1994; Williams, 1997) to explain individual language learning processes in context. The learnt language in our case is German, but we are hoping that our findings will be applicable to the learning and teaching of other languages. The goal of the Mocha project is to build a student model which is informed by current thinking in SLA. The dynamic complexities of second language learning processes for individual learners in the context of computational modeling in CALL necessitates theoretical, analytical framework, which on the one hand is capable of considering the evolving complexity of variables that influence this learning process and on the other is sufficiently formal—in a mathematical sense—to provide the basis for a computational implementation. This is the reason why we are looking to DST to inform our analysis and modeling of individual language learning processes over time.

As has been outlined above, this makes it impossible to employ modeling techniques which rely on the anticipation of errors (bug library and model tracing). The problem with the large search space when using relaxed constraints makes the adoption of this modeling technique very hard. We are currently experimenting with an approach which borrows from machine learning, i.e., the system will be capable to 'learn' new linguistic information and new modeling rules and then be able to handle unanticipated student output. The linguistic analysis is informed by a formal variant of construction grammar[vi] (Kay, 2002). It is our hypothesis that employing this grammatical formalism—with its exclusive reliance on the construction as the unit of analysis and unification to explain the relation of more or less grammatically and lexically specified constructions—will help us to avoid limitations of modeling methods which either rely on error anticipation or on constraint relaxation by identifying well-formed and ill-formed constructions without attempting to anticipate either and by unifying learner constructions which are already known to the learner model. This approach to construction grammar-based parsing is currently only at the stage of conceptualization since the necessary formal approaches to construction grammar are relatively recent and hardly any implementations have been documented in the literature to date.

A formal approach to grammar is, of course, a necessary prerequisite for its computational implementation. However, it is the fact that construction grammar is usage-based and therefore also considers pragmatic features of constructions in addition to morpho-syntactic and semantic features which led to its successful application in first language acquisition research (e.g., Tomasello, 2003). Attempts have also been made to employ it for linguistic analyses in SLA (e.g., Haberzettl, 2007). Tomasello (2003), for example, was able to show that young children acquire their first language by repeating holophrases first—short constructions which consist of fixed lexical material and have to be used with the same meaning and context. Later they manipulate acquired constructions and produce item-based constructions, in which clearly defined parts of the construction are substituted with other suitable items, to then arrive at the level of abstract construction: "a form-meaning pair (F, M) where F is a set of conditions on syntactic and phonological form and M is a set of conditions on meaning and use" (Lakoff 1987, p. 467; quoted in Fischer & Stefanowitsch, 2007, p. 5). Constructions are not just

manipulated over time, but they can also be merged or blended. It is our hypothesis that this acquisition order from more concrete to more abstract constructions can also be applied to SLA and that the usage-based construction grammar is a useful tool for textual analysis under the DST approach. But how can complexity of the language learning system be captured for a student model?

Such complex, dynamic systems, then, can be described by separating data points—the value of variable at a selected system state—and plotting these values in a time series graph: $Y = \{Y_t : t \in T\}$. If the lag time between neighboring data points is identical, then the time series can easily be converted to phase space. Here each data point is seen as a function of a previous data point: $f(x_{t+1}) = a \cdot x_t$. In other words, instead of plotting data points over time, they are plotted 'against each other': $\{(x_t, x_{t+1}), (x_{t+1}, x_{t+2}), (x_{t+2}, x_{t+3}), (x_{t+3}, x_{t+4}), ...\}$. This basically means in our context that the language learning process is plotted in such a way that a language learning event in the past is seen as the starting point or basis for the current language learning event or—to illustrate the predictive power of the student model—the current or past language learning events are understood to be the basis for a language learning event in the immediate future. So, for example, if we have plotted all constructions—by labeling each construction with a numeric identifier first—that a student in an elementary language class has used in the first four weeks, we would find a number of constructions with noun phrases in the nominative or accusative case. We could then make predictions about how to facilitate the learning of indirect objects and their dative-marking, considering this student's understanding of case-marking in the context of her understanding of grammatical phenomena such as word order, verb conjugation as well as based on the evidence the model has collected from the student's text on individual characteristics such as willingness to communicate (based on the frequency and length of her textual contributions relative to her peers) and her willingness to take risks with structural text elements (diversity of lexical and grammatical material identified thus far).

The question which arises is what variable(s) should be measured for plotting at each data point. Our goal of an integrative and balanced view of SLA prevents us from selecting one quantifiable variable and from ignoring the context by trying to eliminate all other variables and their interaction with each other over time. On the other hand, we cannot afford to plot a multitude of variables over time. Our phase space would have as many dimensions as we have variables. The mathematics of multidimensional spaces is complicated and certainly beyond the grasps of this author; some of it is still not known or has not been proven yet. We therefore decided to measure one variable which shows traces of all other linguistic, contextual and individual variables which might play a role: text. We view text as a product (the text we analyze) and as a window onto text as process (the production of text which also reflects the learning process to a large extent). Text is, of course, a very complex variable and it shows traits of a multitude of text-external variables such as individual differences and instructional variables. In other words, by capturing and plotting a student's language use over time in form and their use of constructions in text, we are using the complex variable text as a window on the

complex and dynamic learning process.

We are currently investigating the possibility of measuring the discourse complexity of learner utterances and plotting these over time to get a good approximation of the individual learning process. We selected utterances because they are the smallest linguistic sign with a communicative meaning and because we assume that they incorporate the influence of the dynamic variables that are relevant to this language learning event. In order to measure the complexity, we conduct a construction grammar analysis of each utterance and estimate the level of abstraction for each construction by calculating the frequency of the variants of that construction which the student produced before. Basically, we are attempting to get an approximation of the entropy in the text—how little or often constructions get repeated in text over time and how significantly one construction differs from another—with respect to the construction in question because we can then assume that text entropy and text complexity are proportional. Different instantiations of abstract constructions clearly result in a higher text entropy and are assumed to be an indication of a higher level of the complexity of the learner text. At the other end of the continuum, holophrase constructions which are identical to input material the student was likely to have seen are assumed to have a very low level of discourse complexity. Traditionally, this kind of complexity would have been described informally as the range of vocabulary and the range of grammatical constructions. The different complexity levels are then plotted in a phase space. This graph will give us some indication of what constructions learners used at different times in their language learning process and how they varied in discourse complexity relative to one another.

We have started analyzing written utterances students produced at various stages of a one-semester online course. We are developing our analytical tools—computational German construction grammar—concurrently with our data sets of language learning processes. Using this approach, we intend to model individual cases of complexity and use of learner language first. Also, we intend to examine later whether these individual differences have places of overlap and, if they exist, how they can be used to improve the predictive power of the model.

CONCLUSION

Information from such a student model can be used in diagnostic testing. It will provide a more holistic, balanced picture of the complexity of utterances the learner is able to produce in L2. Having some information about the complexity level an individual learner is at will also enable the system to provide better corrective, error-contingent feedback because the probability of mapping a well-formed and intended construction onto the learner's construction or utterance is much higher. Similarly, the probability of the system being able to suggest a suitable learning object based on a good evaluation of the learner is much higher if the system has information on the prior language learning path and an identification of strengths, weaknesses, and learning preferences.

As research at the intersection of DST, Construction Grammar, SLA, and ICALL is very

new, many unanswered questions remain, and many claims await empirical testing. However, the philosophical and methodological slant of DST as well as Construction Grammar, their mathematical foundations, and their integrative nature hold great promise for further progress in student modeling in ICALL and, probably more importantly, for an improved, more comprehensive understanding of SLA processes.

REFERENCES

Antos, G. (1982). *Grundlagen einer Theorie des Formulierens. Textherstellung in geschriebener und gesprochener Sprache*. Tübingen: Niemeyer.

Burton, R. R., & Brown, J. S. (1982). An investigation of computer coaching for informal learning activities. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 79-88). London: Academic Press.

Cerri, S. A., Cheli, E., & McIntyre, A. (1992). Nobile: Object-based user model acquisition for second language learning. In M. L. Swartz & M. Yazdani (Eds.), *Intelligent tutoring systems for foreign language learning: The bridge to international communication* (pp. 171-190). Berlin: Springer Verlag.

Corder, S. P. (1974). The significance of learners' errors. In J. C. Richards (Ed.), *Error Analysis: Perspectives on second language acquisition* (pp. 19-27). London: Longman.

Corder, S. P. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

de Bot, K., Lowie, W., & Verspoor, M. (2005). Dynamic Systems Theory and Applied Linguistics: The Ultimate "so what"? *International Journal of Applied Linguistics*, *15*(1), 116-118.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7-21.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ; London: Lawrence Erlbaum Associates.

Ellis, N. C., & Larsen-Freeman, D. (Eds.). (2006). *Language emergence - Implications for Applied Linguistics. Special Issue of Applied Linguistics (27/4)*. Oxford: Oxford University Press.

Elsom-Cook, M. (1993). Student modeling in intelligent tutoring systems. *Artificial Intelligence Review*, *7*(3-4), 227-240.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental

concepts in SLA research. *The Modern Language Journal*, *81*(3), 285-300.

Fischer, K., & Stefanowitsch, A. (2007). Konstruktionsgrammatik: Ein Überblick. In K. Fischer & A. Stefanowitsch (Eds.), *Konstruktionsgrammatik. Von der Anwendung zur Theorie* (pp. 3-17). Tübingen: Stauffenberg Verlag.

Gisolfi, A., Dattolo, A., & Balzano, W. (1992). A Fuzzy Approach to student modeling. *Computers and Education*, *19*(4), 329-334.

Gleick, J. (1987). *Chaos: Making a new science*. New York, N.Y.: Viking.

Haberzettl, S. (2007). Konstruktionen im Zweitsprachenerwerb. In K. Fischer & A. Stefanowitsch (Eds.), *Konstruktionsgrammatik. Von der Anwendung zur Theorie* (pp. 55-77). Tübingen: Stauffenberg Verlag.

Haggis, T. (2005). *'Knowledge Must Be Contextual': Some possible implications of Complexity and Dynamic Systems Theories for educational research.* Paper presented at the Complexity, Science and Society Conference, Liverpool.

Heift, T. (1998). Designed Intelligence: A Language Teacher Model. Unpublished PhD Thesis, Simon Fraser University, Burnaby.

Heift, T., & Nicholson, D. (2001). Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education*, *12*(4), 310-325.

Heift, T., & Schulze, M. (2007). *Intelligence and errors in CALL. Parsers and pedagogues*. New York: Routledge.

Heinecke, J., Kunze, J., Menzel, W., & Schröder, I. (1998). Eliminative parsing with graded constraints. In *Proceedings of Coling-ACL'98* (pp. 526-530).

Herdina, P., & Jessner, U. (2002). *A dynamic model of multilingualism: Perspectives of change in psycholinguistics*. Clevedon; Buffalo; Toronto: Multilingual Matters.

Hollich, G. J. (2000). Mechanisms of word learning: A computational model. Unpublished Dissertation/Thesis, http://www.il.proquest.com/umi, Univ Microfilms International, US.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., et al. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*(3), v-123.

Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Oxford: Blackwell.

Kay, P. (2002). An informal sketch of a formal architecture for construction grammar.

Grammars, 5(1), 1-19.

Klein, W. (1986). *Second language acquisition*. Cambridge: CUP.

Klein, W., & Dittmar, N. (1979). *Developing grammars. The acquisition of German syntax by foreign workers*. Berlin: Springer Verlag.

Klein, W., & Perdue, C. (1992). *Utterance structure (Developing grammars again).* Amsterdam/Philadelphia: Benjamins.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon.

Lado, R. (1957). *Linguistics across cultures: Applied Linguistics for language teachers*. Ann Arbor: The University of Michigan Press.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind.* Chicago: University of Chicago Press.

Larsen-Freeman, D. (1997). Chaos/Complexity Science and second language acquisition. *Applied Linguistics, 18*(2), 141-165.

Larsen-Freeman, D. (2000). Second language acquisition and Applied Linguistics. *Annual Review of Applied Linguistics, 20*, 165-181.

Larsen-Freeman, D. (2003). *Teaching language: From grammar to grammaring*. Southbank, Victoria: Thomson/Heinle.

Levy, M. (1997). *Computer-assisted language learning: Context and conceptualisation.* Oxford: Oxford University Press.

Lorenz, E. N. (1993). *The Essence of chaos*. Seattle: University of Washington Press.

Mabbott, A., & Bull, S. (2004). Alternative views on knowledge: Presentation of open learner models. In J. C. Lester, R. M. Vicari & F. Paraguacu (Eds.), *Intelligent Tutoring Systems: 7th International Conference* (pp. 689-698). Berlin: Springer-Verlag.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology, 49*, 199-227.

McCalla, G. I., Vassileva, J., Greer, J. E., & Bull, S. (2000). Active learner modelling. In G. Gauthier, C. Frasson & K. VanLehn (Eds.), *Intelligent Tutoring Systems. 5th International Conference, ITS 2000, Montréal, Canada, June 2000, Proceedings* (pp. 53-62). Berlin: Springer Verlag.

Menzel, W. (1988). Error diagnosing and selection in a training system for second language learning. In *Proceedings of the Twelfth International Conference on*

Computational Linguistics (pp. 414-419).

Menzel, W. (1992a). Constraint-based diagnosis of grammatical faults. In J. Thompson & C. Zähner (Eds.), *Proceedings of the ICALL Workshop, UMIST, September 1991* (pp. 89-101). Hull: University of Hull, CTI Centre for Modern Languages.

Menzel, W. (1992b). *Modellbasierte Fehlerdiagnose in Sprachlernsystemen*. Tübingen: Niemeyer.

Menzel, W., & Schröder, I. (1998). Constraint-based diagnosis for intelligent language tutoring systems. In *Proceedings of the IT&KNOWS Conference at IFIP'98 Congress* (pp. 484-497). Wien/Budapest.

Milne, S., Shiu, E., & Cook, J. (1996). Development of a model of user attributes and its implementation within an adaptive tutoring system. *User Modeling and User-Adapted Interaction, 6*, 303-335.

Mislevy, R. J., & Gitomer, D. H. (1996). The Role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction, 5*, 253-282.

Mitrovic, A. (1998). Experiences in implementing constraint-based modeling in *SQL-Tutor*. In B. P. Goettl, H. M. Halff, C. L. Redfield & V. J. Shute (Eds.), *Intelligent Tutoring Systems. 4th International Conference, ITS 1998, San Antonio, Texas, USA, August 1998, Proceedings* (pp. 414-423).

Mitrovic, A., Djordjevic-Kajan, S., & Stoimenov, L. (1996). INSTRUCT: Modeling students by asking questions. *User Modeling and User-Adapted Interaction, 6*, 273-302.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education, 10*, 98-129.

Ohlsson, S. (1992). Constraint-based student modeling. *Journal of Artificial Intelligence in Education, 3*(4), 429-447.

Reyes, R. L. (1998). A domain theory extension of a student modeling system for Pascal programming. In B. P. Goettl, H. M. Halff, C. L. Redfield & V. J. Shute (Eds.), *Intelligent Tutoring Systems. 4th International Conference, ITS 1998, San Antonio, Texas, USA, August 1998, Proceedings* (pp. 324-333).

Schachter, J. (1974). An error in Error Analysis. *Language Learning, 27*, 205-214.

Schönefeld, D. (2006). Constructions [Electronic Version]. *Constructions, SV 1*, from www.constructions-online.de

Schröder, I. (2002). Natural language parsing with graded constraints. Unpublished PhD

Thesis, University of Hamburg, Hamburg.

Schulze, M. (1998). Teaching grammar - learning grammar: Aspects of second language acquisition. *CALL, 11*(2), 215-228.

Schulze, M. (1999). From the developer to the learner: Computing grammar - learning grammar. *ReCall, 11*(1), 117-124.

Schulze, M. (2001). Textana - Grammar and grammar checking in parser-based CALL. Unpublished PhD Thesis, UMIST, Manchester.

Selinker, L. (1974). Interlanguage. In J. C. Richards (Ed.), *Error Analysis: Perspectives on second language acquisition* (pp. 31-54). London: Longman.

Selinker, L. (1992). *Rediscovering Interlanguage*. London: Longman.

Swain, M. (1985). Communicative competence: Some roles of Comprehensible Input and Comprehensible Output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley: Newbury House.

Tasso, C., Fum, D., & Giangrandi, P. (1992). The Use of explanation-based learning for modelling student behavior in foreign language tutoring. In M. L. Swartz & M. Yazdani (Eds.), *Intelligent tutoring systems for foreign language learning: The bridge to international communication* (pp. 151-170). Berlin: Springer Verlag.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, Mass. ; London: Harvard University Press.

van Geert, P. (1994). *Dynamic Systems of Development: Change between Complexity and Chaos*. Harvester Wheatsheaf, Hertfordshire, HP2 7EZ: England.

van Geert, P. (1997). Nonlinear Dynamics and the Explanation of Mental and Behavioral Development. *Journal of Mind and Behavior, 18*(2-3), 269-290.

van Geert, P. (1999). *Vygotsky's Dynamic Systems*. Taylor & Frances/Routledge, Florence, KY, US, [URL:http: //www.routledge.com].

van Geert, P. (2000). The Dynamics of General Developmental Mechanisms: From Piaget and Vygotsky to Dynamic Systems Models. *Current Directions in Psychological Science, 9*(2), 64-68.

van Geert, P., Verhoeven, L., & van Balkom, H. (2004). A dynamic systems approach to diagnostic measurement of SLI. In *Classification of developmental language disorders: Theoretical issues and clinical implications.* (pp. 327-348). Lawrence Erlbaum Associates, Publishers, Mahwah, NJ: US.

Vandeventer, A. (2001). Creating a Grammar Checker for CALL by Constraint

Relaxation: A Feasibility Study. *ReCall, 13*(1), 110-120.

Villano, M. (1992). Probabilistic Student Models: Bayesian Belief Networks and Knowledge Space Theory. In C. Frasson, G. Gauthier & G. I. McCalla (Eds.), *Intelligent Tutoring Systems. Second International Conference, ITS'96, Montréal, Canada, June 1992, Proceedings* (pp. 491-498). Berlin: Springer Verlag.

Williams, G. P. (1997). *Chaos Theory Tamed*. Washington, D.C.: Joseph Henry Press.

---

[i] Different theoretical approaches are subsumed here (for reasons of stylistic convenience rather than because they are identical): Dynamic Systems Theory, Chaos Theory, Complexity Theory, Emergentism (see e.g., Gleick, 1987; Lorenz, 1993; van Geert, 1994; Williams, 1997).

[ii] A multitude of variables influence the path of this process and potentially 'decide' about success or failure. It is the interaction of these variables which contribute to the "emergence of language" (MacWhinney, 1998).

[iii] For a more detailed discussion of different modeling techniques and student models in CALL see Heift and Schulze (2007).

[iv] Here I follow Antos (1982) who argued that although text production might not be a problem solving process, it is fruitful in Applied Linguistics research to depict it as such.

[v] Principal investigator: Mathias Schulze; Co-investigator: Trude Heift; The research is supported by the Social Sciences and Humanities Research Council (SSHRC) of Canada; grant number 410-2007-2549

[vi] Construction grammar is an umbrella term for a number of more or less formal approaches which all have in common that they view constructions as the central syntactic unit. For an overview, see (Fischer & Stefanowitsch, 2007; Schönefeld, 2006).

Baralt, M. (2008). Lexical acquisition, awareness, and self-assessment through computer-mediated interaction: The effects of modality and dyad type. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 167-193). Ames, IA: Iowa State University.

Lexical Acquisition, Awareness, and Self-Assessment through Computer-Mediated Interaction:
The Effects of Modality and Dyad Type

Melissa Baralt
Georgetown University

It has recently been demonstrated that interaction within the CMC (computer-mediated communication) modality can provide many of the same benefits as face-to-face (FTF) interaction (De la Fuente, 2003; Smith, 2004, 2005; Shekary & Tahririian, 2006, Sachs & Suh, 2007). One of the main premises behind the use of CMC tasks is that any development acquired through CMC might eventually be transferred to the oral mode. In addition, these chat dialogues can be saved and reviewed for later analysis, making CMC chat a unique tool for L2 conversational practice and assessment. This study is twofold: the first experiment examines the potential for lexical acquisition by beginning-level learners in CMC as compared to face-to-face (FTF). Furthermore, different dyad-partner proficiency levels were used, as Iwashita (2001) has suggested that mixed-proficiency dyads elicit more instances of negotiation and recasts. Results indicate that beginning-level learners did significantly better in the CMC mode than the FTF mode on oral and written production tests. Type of dyad (or the proficiency level of the beginning learner's partner) did not have an effect on learning. Experiment 2 describes sessions between four beginning-level learners and the researcher in regards to their saved CMC chat files from Experiment 1. An analysis of their saved and stored "conversations" revealed that learners were able to identify errors, recognize reasons for instances of non-understanding that took place with their CMC partner, and spot, as well as correct, problems in their interlanguage. It is argued that the benefits of saving iChat conversations as a "record-keeping" of learners' interactional abilities include placing assessment in the hands of the learners and providing personal records for learners to monitor their progress over time.

INTRODUCTION

The Interactionalist approach to second language acquisition posits that when learners have the opportunity to negotiate for meaning with their interlocutor, second language acquisition can be facilitated (Long 1996). While engaging in conversation, learners can receive negative feedback that allows them to reformulate and develop their interlanguage. Negative feedback can take form as recasts, explicit grammatical correction, questioning, confirmation checks, or indications of non-understanding. According to Long, it might be that both conversational environments and the internal processing of the learner are important for the negotiation of meaning that pushes second

language development. Feedback that is provided in negotiation work is precisely what may facilitate learners' attention being drawn to areas of language that they need to focus on for L2 development.

While negotiating for meaning within conversational interaction, learners also have the opportunity to produce output (Swain 1985, 1995). Swain posits that it is the need to produce language that causes learners to think about the interlanguage, and that input might not be enough for certain aspects of L2 acquisition. According to her output hypothesis, learners engaging in interaction are 'pushed' to produce comprehensible input so that they can be understood. She suggests that pushed output enhances fluency and causes the learner to test hypotheses about his or her metalinguistic knowledge (Swain 2005). This type of output then prompts recasts from a Non-native-speaking (NNS) or Native-speaking (NS) interlocutor in conversation, which can lead to episodes of negotiation for meaning. This paper draws upon this line of research to describe research investigating the use of computer-mediated communication for L2 acquisition.

PREVIOUS RESEARCH

Empirical studies Operationalizing Interaction and Pushed Output

Several studies have empirically demonstrated that interaction is beneficial for L2 development, lending support to Long's hypothesis (Ellis, Tanaka and Yamazaki, 1994; Gass & Varonis, 1994; Mackey, 1999; Pica, Yong & Doughty, 1987). Others have also looked specifically at the potential for *learning* via interaction in the face-to-face (FTF) modality, lending support to both the Interaction and Pushed Output Hypotheses. For example, de la Fuente (2002) demonstrated the importance for both negotiation for meaning and produced output. She found that a combination of negotiated interaction and pushed output was the only treatment that promoted both receptive and productive acquisition of words. Similarly, Ellis and He (1999) examined the acquisition of lexical items, and found that the modified output group surpassed the two input groups in comprehension and vocabulary gain scores. These studies, both in the FTF oral mode, showed that when learners have the opportunity to negotiate for meaning and produce and modify their output, language learning can take place.

Interaction in Computer-Mediated Communication

Since the late 1990's, researchers have attempted to extend the potentiality for language learning from the FTF to the computer-mediated communication (CMC) modality. This has been in large part due to the increased popularity of technology in the second language classroom, especially due to the ever-growing need for distance learning programs. Synchronous CMC is a virtual, real-time conversation that takes place across a computer network such as the Internet. The premise behind incorporating CMC chat into the SLA classroom is that it provides students with the opportunity to practice and interact with each other in their second language, perhaps as extra practice outside of the classroom, for projects, or for distance learning. The assumption is that synchronous

electronic chat is analogous to oral or face-to-face chat. Pellettieri (2000, p. 59) clearly states this postulation: "…because synchronous [CMC] chatting bears a striking resemblance to oral interaction, it seems logical to assume that language practice through [CMC] will reap some of the same benefits for second language development as practice through oral interaction." In her study, Pellettieri found that in tasks conducted in NBC (what she calls network-based communication), negotiation for meaning occurred, and in fact, learners' "patterns of interaction looked very much like those seen in NNS oral conversation" (2000, p. 70). She called for more research that provides empirical support for this assumption. Several researchers have attempted to do that by examining the interactive discourse in CMC (Beauvois, 1992; Blake, 2000; Chun, 1994; Darhowever, 2002; Fernández-García and Martínez-Arbelaiz, 2002; Kötter, 2003; Pelletieri, 2000; Smith; 2003; Toyoda and Harrison, 2002; Tudini, 2003; Warner, 2003; Xie, 2002); teacher strategies in CMC (Meskill, 2005); synchronous CMC as compared to asynchronous CMC (Pérez, 2003); CMC chat compared to FTF chat (Abrams 2003; Böhlke, 2003; Fitz, 2006; Kern, 1995; Lai and Zhao, 2006; Salaberry, 2000; Sykes, 2005; Vandergriff, 2006; Warschauer, 1996); socialization processes in the CMC modality `(Sengupta, 2001; Shin, 2006); and computerized written chat rooms as compared to computerized oral chat rooms (Jepson, 2005). Some researchers have also reported that the CMC modality reduces learner anxiety (Kern, 1995) and can help to equalize student participation (Xie, 2002).

Since 2003, five empirical studies have shown that language learning can take place through CMC, (De la Fuente, 2003; Smith, 2004; 2005; Shekary & Tahririan, 2006; Sachs and Suh, 2007). All of these studies lend empirical support that negotiation for meaning in synchronous CMC chat can facilitate L2 learning of linguistic items. Of these five studies, only de la Fuente (2003) examined the differences between CMC and FTF L2 learning. While some studies have compared interactive moves in CMC with those of FTF (Abrams 2003; Böhlke, 2003; Kern, 1995; Lai and Zhao, 2006; Salaberry, 2000; Sykes, 2005; Warschauer, 1996), de la Fuente (2003) is to date the only study that has attempted to compare how learning is achieved in CMC versus FTF and in what ways it might be different. In her study, de la Fuente (2003) randomly assigned participants from three second-semester classes of Spanish into two groups: oral FTF interaction and virtual chat (CMC) to assess how well each group learned the Spanish names of 14 fruit, vegetable and seafood words. She employed a pretest-posttest-delayed posttest design. For the productive scores, students spoke (for oral measurements) and wrote (for written measurements) the names of the food items. For receptive scores, students listened to names and tried to say their names in English (oral) and then were required to translate a list of words into English in writing. During the treatment, participants were given an information gap task in which they had to assign to each other certain food items that they needed their partner to retrieve from the market. De la Fuente found that both groups – the oral interaction group and the virtual chat group – had receptive and productive gains in the acquisition of the L2 vocabulary items. Though the oral interaction group outperformed the virtual chat group in written tests, differences were not significant. De la Fuente had hypothesized that the CMC group would outperform the FTF group

because of previous claims that CMC may prompt students to pay more attention to the targeted forms (given that it is slower and students can visually see their own and their partner's output). She found however that the oral interaction group had higher productive acquisition one day and one week after the treatment, while the virtual chat group did not. De la Fuente concludes that type of medium does not affect learning, although face-to-face interaction might be more beneficial than CMC for short-term oral productive acquisition. She makes the claim that CMC is as effective as FTF, but it is not necessarily better, especially in terms of oral production.

One limitation of de la Fuente's study is the extremely controlled time limitation allotted to episodes of negotiation for meaning. In the virtual chat group, participants were allowed two minutes to negotiate the meanings of the vocabulary items, while the oral interaction group was given only one minute. De la Fuente does not provide a justification for this control on time. In fact, in looking at her data, we see that the participants appear rushed, and moved on to the next item before finishing the first item (and hence before successfully achieving an understanding of its meaning):

> Example 5, Pair 1, Group 2, Day 1 (Taken from de la Fuente 2003 pg. 72):
>
> sth5> *necesito frambuesas* [I need cranberries]
> sth5> *son rojas y pequenos* [they are red and small]
> ag20> *en ensaladas o no?* [in salads or not?]
> sth5> *no, es una fruto* [no, it is a fruit]
> sth5> *no es caliente* [it's not hot]
> ag20> *es un tipo como cerezas?* [is it a type like cherries?]
> sth5> *no se* [I don't know]
> sth5> *es un poco raro* [it is a bit strange]
> ag20> *no es importante* [it is not important]
> ag20> *pues, no se …* [well, I do not know…]
> ag20> *vamos a tratar las dos y despues eso* [let's try two and then this one]
> sth5> *bien* [ok]
> sth5> *necesito ciruelas* [I need plums]

The dialogue above shows that the two participants in pair 1 do not achieve a successful negotiation for meaning before they decide to go on to the next item. Because of the time restraint, participant ag20 suggests that they try the next one and then come back to this one. De la Fuente does not report if they did return to this item or were ever successful in getting its meaning across. Perhaps if the participants in her study had had more time, they may have been able to negotiate the meaning of the word together. Another limitation to her study is that treatment sessions were held on two consecutive days, with no debriefing questionnaire to ensure that participants did not go back and look up any of the items, or that they had outside influence on top of the treatment itself that may have affected their posttests.

Effects of Dyad Partner Proficiency Level

One other characteristic of de la Fuente's study is she only utilized NNS-NNS dyads of intermediate learners. In fact, most of the studies that have examined the learning potential in the CMC modality (De la Fuente, 2003; Shekary & Tahririan, 2006; Smith, 2004, 2005) use only NNS-NNS dyads of intermediate learners, all at the same proficiency level. It might be worthwhile to explore mixed-proficiency dyads, as it has been empirically demonstrated that mixed-proficiency dyads provide more interactional moves than same level dyads. Iwashita (2001) examined interaction and modified output between 24 learners of Japanese. Participants were divided into three groups of dyads: High-High, Low-Low, and High-Low. Each dyad performed one jigsaw and two information gap tasks in the oral mode. Iwashita coded their interaction for c-units, or utterances that contain communicative value. He found that low proficiency learners in mixed proficiency dyads (High-Low) modified output more than low proficiency learners in the same proficiency dyads (Low-Low) – an important finding when considering what type of dyad should be employed in designing tasks. Iwashita's study was in the oral mode. Whether differential effects for dyad-partner proficiency level applies to the CMC mode has yet to be explored. Given that the existing literature has established a connection between interactional negotiation of meaning and L2 acquisition in both the FTF and CMC modes, it is justifiable to ask whether dyad-partner proficiency level also has an effect on learners' L2 learning in either the FTF or CMC modalities.

Prior Claims on Benefits of CMC as Compared to FTF

Studies that examined discourse in the CMC mode and compared it to FTF have found more beneficial results for the CMC mode, such as higher quantity of language produced in CMC than FTF (Abrams, 2003), more equalized participation in CMC than FTF (Böhlke, 2003), observation of change in morphosyntactic development more identifiable in CMC than FTF (Salaberry, 2000), and more lexically and syntactically complex language produced in the CMC mode than the FTF mode (Warschauer, 1996). The findings of these studies imply a possible advantage for practicing language in the CMC mode, at least in the beginning stages of L2 development. More research, with more robust designs, is needed that shows strong empirical evidence for the learning potential in the CMC mode as compared to the FTF mode. It could be that FTF and CMC differentially promote oral and/or written acquisition, but not enough evidence comparing the two is available to be conclusive. Empirical research is needed that investigates a) the relative effects of CMC versus FTF communication on L2 learning, b) the differences present in discourse of CMC versus FTF and how they might differentially affect learning, and c) how we as researchers can capitalize on the differential benefits of CMC as compared to FTF, if such benefits do exist.

One of these benefits could be the use of CMC as an assessment tool. Assessment here refers to the learner's retrospective assessment of his or her recorded linguistic production. CMC chat can be conducted in the written modality, and since chat logs can

be saved, it would be useful to see what participants do with their own saved 'conversations' in a follow-up session with the investigator.

PURPOSE AND RESEARCH QUESTIONS

To this end, the present study is divided into two experiments. Experiment 1 compares the effects of modality (CMC or FTF) and dyad partner proficiency level (Beginning-Beginning, Beginning-Advanced, Beginning-Native Speaker) on the beginning level learner's acquisition of lexical items. In Experiment 2, saved iChat conversations from Experiment 1 were used with 4 participants to see how such files may be used for assessment purposes. The following research questions guide this study:

Experiment 1

1. Does modality (CMC versus FTF) have an effect on beginning learners' acquisition of lexical items, as measured by a) oral production, b) written reception, and c) written production tests?

2. Does the beginning learner's partner proficiency level (Beginner, Advanced, or Native Speaker) have an effect on the beginning learner's acquisition of lexical items, as measured by a) oral production, b) written reception, and c) written production tests? If so, which mixed-proficiency dyad facilitates learning the best?

Experiment 2

3. What do learners do in going back and reviewing their saved iChat files with the instructor?

4. Can saved chat logs be used as self-diagnostic tools for assessment of L2 production?

EXPERIMENT 1

Method

Participants

The Participants in this study originated from a group of 115 beginning and advanced learners of Spanish at a large public university in the United States. Specifically, 45 students were from Advanced Spanish I or Intensive Advanced Spanish I or II levels, while 70 students were in the Beginning Spanish, Beginning Spanish II, or Intensive Beginning Spanish course levels. All were between the ages of 18 and 25, and were all native speakers of English. From the 115 students who initially signed up for the study, participants were eliminated for 1) not fully finishing the task, 2) not following directions during the task (as revealed by the chat dialogue or the video-taped FTF session) or 3) not attending all sessions. In the end, a total of 54 Spanish learners, 42 beginning and 12

Table 1.Groups used according to modality and dyad partner-proficiency level

Face-to-Face (FTF)			Computer-Mediated-Communication (CMC)		
Beginning with Beginning	Beginning with Advanced	Beginning with Native Speaker	Beginning with Beginning	Beginning with Advanced	Beginning with Native Speaker
5 dyads (10 total)	7 dyads (18 total)	9 dyads (18 total)	4 dyads (8 total)	5 dyads (10 total)	4 dyads (8 total)

advanced, comprised the participant group that could be analyzed for this study. Ten native speakers, all from Spain, also participated. Four of the native speakers performed the activity twice, for a total of 15 dyads with learners and native speakers. This resulted in 34 dyadic pairs being analyzed for this study, as seen in Table 1.

Target item

The targeted linguistic items employed in this study were Spanish lexical items. These words were types of chores that one typically does around the house. A chore then could be a verb or an item that needs to be cleaned. From an original list of 22 chores on the pretest, 15 words for which the participants showed no prior knowledge (excluding the native speakers) were chosen to serve as the lexical items for the treatment. The items were: *fregadero* [sink], *regar* [to water], *lavavajillas* [dishwasher], *verja* [fence], *podar* [to trim], *desatascar* [to unclog], *váter* [toilet], *triturador* [food disposal], *telarañas* [spiderwebs], *pulir* [to polish], *aspirar* [to vacuum], *sacudir* [to beat, for example a rug], *trastero* [storage room], *deshollinar* [to chimney-sweep], and *tender* [to hang]. Several native speakers of Spanish, all from different countries, were consulted to get an idea of what types of chore names exist. While it is important to consider dialectal differences as well as 'standard' lexicons for the purpose of pedagogy, in the interest of consistency, the researcher decided to choose chore names that are commonly used in Spain.

As some of the chores were verbs, the task was formatted so that participants would not focus on verb conjugation but rather just tell their partner the name of the chore and negotiate its meaning. Chore assignment was therefore prompted with the script: "Tú tienes que…[say chore item here]" [*You have to … [say chore item here]*]. For example, "*Tú tienes que… limpiar el fregadero,*" with "fregadero [sink]" serving as the lexical item to be negotiated.

Materials

Apple Inc.'s iChat software, version 3.1.9., was used for the CMC group. iChat runs on an instant message framework, and is used on any Mac OS X operating system. In iChat, messages from the participants' partners appear as a callout bubble on either the left or right side of a dialogue box that participants see on the computer. Pre-selected icons indicate who sent the message, as each icon appears in message callout bubbles. Colors of the text, the callout bubble background, and font types can all be chosen by the

participants. Each chat conversation can be saved as an iChat file, and later copied and pasted into a word file.

Task

An information gap task was used for this study. Since students had just returned from spring break, they were told that it was now time to do some "Spring Cleaning" to correspond with their vacation. For the Beginner-Beginner (Beg-Beg) and Beginner-Advanced (Beg-Adv) dyads, participants were asked to assign their partner seven chores, and then switch roles so that the other person could assign chores. Participants then repeated the activity, once more switching roles, so that in the end each participant had been able to assign and be given each lexical item. The task was completed on the computer in iChat for the CMC group, and in person for the FTF group. In both modalities, dyads comprised of beginning-level learners and native speakers had slightly different instructions. Because the native speaker would obviously understand the chore being assigned to him/her, the beginning level learner had to instead describe the chore to make the native speaker guess it. The beginning-level learner and his or her native speaker partner then switched roles, as the Beg-Beg and Beg-Adv groups did. All participants in the FTF group were videotaped with a digital camcorder so that conversations could be analyzed, while those in the CMC group had their iChat conversations saved as computer files for further analysis.

Testing instruments

To measure lexical acquisition, a production test (both oral and written) and a reception test (written) were used. For both the oral and written production tests, participants were given a sheet of paper with pictures of all of the chores. Participants first had to say the names of each chore aloud in Spanish (oral production), and then write the name of each chore in Spanish (written production). Answers provided in the oral mode were recorded with the software Audio High Jack Pro on the computer, and saved as MP3 files. For the written reception test, participants were given a list of the chores in Spanish and asked to simply write the English equivalents if they knew them.

Procedure

As Figure 1 demonstrates, this study consisted of a pretest-treatment-immediate posttest-delayed posttest design.

Two weeks before the experiment, participants were pretested on a list of 22 potential lexical items. After the pretest, participants signed up for treatment session times. Hence, dyad partners were paired together randomly depending on which days they were available. As this study examined the potential acquisition gains of the beginning level learner, dyads were established that paired a beginning-level learner with either 1) another beginning-level learner, 2) an advanced learner, or 3) a native speaker. Over a span of four days, participants came into the lab or into a classroom to take part in the treatment session. They performed the task either in person (FTF) with their partner, or via iChat on the computer (CMC). Before starting the task, participants were given three

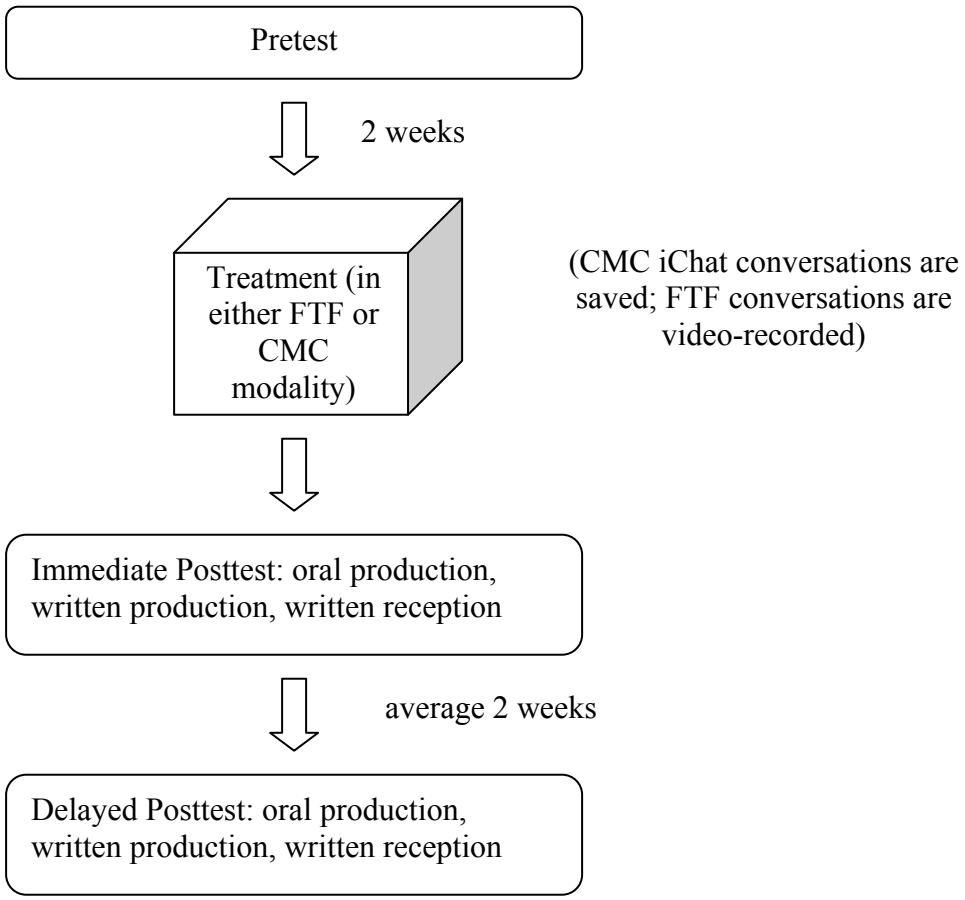


Figure 1. Study design

minutes to talk with their partners either in person (FTF) or on the computer (CMC) as a warm-up session. Immediately after the treatment, participants took an oral production test, a written production test, and then a written test of receptive knowledge. The production tests preceded the receptive test to eliminate the possibility of giving students cues about the words they would need to produce. Participants then filled out a questionnaire in which they were asked for their opinions and comments on the task they had just done with their partner. Approximately two weeks later, all participants (except for the native speakers) came in to take the delayed posttests, which also consisted of an oral production, a written production, and a written reception test[i].

Scoring

Scores for the oral production and written production were derived from the correct oral and written answers provided by the participants. With the written recognition test, scores came from the correct English translation provided for each targeted item. One point was assigned to each correct answer. The maximum potential score for each test was 15.

During the scoring process, all efforts were made to assign correct points only to well-formed oral and written answers. However, some minor deviations from the correct form were accepted as correct answers if they did not indicate lack of the basic lexical meaning. For example, incorrect gender assignment was not considered incorrect (*fregadera* as opposed to the correct *fregadero* [sink], or *trituradora* as opposed to *triturador* [garbage disposal].)[ii]

Results

RQ1: The effects of modality (CMC vs. FTF) on learning

To measure the effects of modality on the beginning learners' acquisition of lexical items, data were submitted to a repeated-measures analysis of variance (ANOVA) using a two between-subject and three within-subject factorial design. Modality (CMC versus FTF) served as the between-subject factor, while Time (Pretest vs. Immediate Posttest vs. Delayed Posttest) served as the within-subject factor. A select-cases filter was implemented on the data, with the condition of 1 representing beginning-level learners. In this way, only the beginning-level learners' data (in relation to their type of dyad) would be analyzed, as the point of this study was to measure their acquisition gains (and not the gains of advanced learners or the native speakers).

Oral production. Mean scores and standard deviations for productive oral acquisition of both the CMC and FTF groups on each test day are provided in Table 2. The results of the ANOVA showed that beginning-level learners did significantly better in the CMC mode than the FTF mode on oral production tests (F (1, 32) = 5.78, p =.022), as seen in Table 3. Post hoc tests revealed that differences between the FTF and CMC

Table 2. Mean scores and standard deviations (SD) for oral production (depending on modality)

	Modality	Mean	Standard Deviation	n
Oral Production	FTF	.000	.000	21
pretest	CMC	.000	.000	13
	Total	.000	.000	34
Oral Production	FTF	2.23	1.75	21
Imm. posttest	CMC	4.34	2.56	13
	Total	3.04	2.30	34
Oral Production	FTF	.907	1.28	21
Delayed posttest	CMC	1.08	1.62	13
	Total	.974	1.39	34

Table 3. ANOVA (between-subject effects for productive oral acquisition)

Source	*df*	*SS*	*MS*	*F*
Intercept	1	65.45	65.45	81.1*
Group (Modality)	1	4.66	4.66	5.78*
Error	32	25.82	.807	

Note: *p ≤ .022

Table 4. Mean scores and standard deviations (SD) for written reception (depending on modality)

	Modality	Mean	Standard Deviation	n
Written Reception Pretest	FTF	.000	.000	21
	CMC	.000	.000	13
	Total	.000	.000	34
Written Reception Imm. posttest	FTF	6.19	3.09	21
	CMC	7.77	3.27	13
	Total	6.79	3.21	34
Written Reception Delayed posttest	FTF	2.81	2.16	21
	CMC	3.54	3.36	13
	Total	3.08	2.66	34

group were significant for changes observed from the pretest to the immediate posttest (p = .007). However, no significant differences were observed between the groups for the immediate posttest to the delayed test.

Written reception. The mean scores and standard deviations for the written receptive task for each group (CMC vs. FTF) are provided in Table 4. To measure acquisition of recognition of the written form of the lexical items in regards to group differences, data were likewise submitted to a 2 x 3 Repeat Measures ANOVA. Though the CMC group had higher gains than the FTF group (7.77 vs. 6.19 respectively on the immediate posttest, and 3.54 vs. 2.81on the delayed posttest), unlike the oral production

Table 5. Mean scores and standard deviations (SD) for written production (depending on modality)

	Modality	Mean	Standard Deviation	n
Written Production Pretest	FTF	.000	.000	21
	CMC	.000	.000	13
	Total	.000	.000	34
Written Production Imm. posttest	FTF	3.00	2.35	21
	CMC	5.92	2.75	13
	Total	4.12	2.86	34
Written Production Delayed posttest	FTF	.524	.814	21
	CMC	.846	.898	13
	Total	.647	.849	34

Table 6. ANOVA (between-subject effects for written production acquisition)

Source	*df*	*SS*	*MS*	*F*
Intercept	1	94.5	94.5	96.0*
Group (Modality)	1	9.39	9.39	9.54*
Error	32	31.5	.985	

Note: *p ≤ .022

measures, no significant differences were found for the main effect of Group (CMC and FTF) for written receptive knowledge. Similarly, no significant interaction between Time and Group was found.

Written production. Mean scores and standard deviations for the oral productive task for the CMC and FTF groups is provided in Table 5. To measure written productive acquisition of lexical items, data were again submitted to a 2 x 3 Repeated Measures ANOVA. Between-subjects tests showed a significant main effect for Group (FTF versus CMC) on test scores F (1, 32) = 9.54, p = .000, as demonstrated in Table 6. Post hoc tests were conducted to further examine the main effect. A significant interaction between Time and Group was also observed F (2, 64) = 10.206, p = .000. Post hoc tests

demonstrated that the differences between the FTF and CMC group were significant for both the immediate posttest and the delayed posttest, with the CMC group showing significantly more gains in written production than the FTF group.

RQ 2: The effects of dyad-partner's level on learning

Research Question 2 addressed whether the beginning learner's partner proficiency level (i.e., type of dyad: Beg-Beg, Beg-Adv, Beg-NS) had an effect on THE beginning learner's acquisition of lexical items. To measure the effect of partner's proficiency level, data were submitted to a repeated-measures ANOVA using a three between-subjects and three within-subjects factorial design. Partner Proficiency Level (Beginning vs. Advanced vs. Native Speaker) served as the between-subjects factor, while Time (Pretest vs. Immediate Posttest vs. Delayed Posttest) served as the within-subjects factor. As stated above, a select-cases filter was put on the data, with the condition of 1 = beginning learner. In this way, only the beginning-level learners' data (in relation to their type of dyad) would be analyzed, as the point of this study was to measure beginning-level learners' acquisition gains and not those of the Advanced or Native Speaker participants.

Oral production. The mean scores and standard deviations for the oral productive acquisition task are provided in Table 7. While beginning learners in all dyad types (Beg-Beg, Beg-Adv, Beg-NS) made gains from the pretest to the posttest, no significant main effect for Partner Proficiency Level was found for any of the measures of Beginning-level learners' oral production on any test. Beginners in the Beg-Adv dyads achieved slightly higher gains than those in the Beg-Beg and Beg-NS dyads on the immediate posttest, however these gains were not significant.

Table 7. Mean scores and standard deviations (SD) for oral production acquisition (depending on dyad type)

	Dyad Type	Mean	Standard Deviation	n
Oral Production Pretest	Beg-Beg	.000	.000	9
	Beg-Adv	.000	.000	11
	Beg-NS	.000	.000	14
	Total	.000	.000	34
Oral Production Imm. posttest	Beg-Beg	2.77	2.59	9
	Beg-Adv	3.64	2.39	11
	Beg-NS	2.74	2.11	14
	Total	3.04	2.30	34
Oral Production Delayed posttest	Beg-Beg	.556	.527	9
	Beg-Adv	1.06	1.64	11
	Beg-NS	1.17	1.59	14
	Total	.974	1.39	34

Table 8. Mean scores and standard deviations (SD) for written reception acquisition (depending on dyad type)

	Dyad Type	Mean	Standard Deviation	n
Written Repection Pretest	Beg-Beg	.000	.000	9
	Beg-Adv	.000	.000	11
	Beg-NS	.000	.000	14
	Total	.000	.000	34
Written Reception Imm. posttest	Beg-Beg	6.78	3.11	9
	Beg-Adv	7.36	2.94	11
	Beg-NS	6.36	3.61	14
	Total	6.79	3.21	34
Written Reception Delayed posttest	Beg-Beg	2.11	1.83	9
	Beg-Adv	3.63	2.42	11
	Beg-NS	3.29	3.22	14
	Total	3.09	2.66	34

Written reception. Once again, mean scores and standard deviations for the written receptive acquisition task are reported in Table 8. As with the oral production results, beginning-level learners in all dyad types achieved gains in written recognition from the pretest to the posttest. However, no significant effect for Partner Proficiency Level on beginners' written reception was found.

Written production. The mean scores and standard deviations for the written productive acquisition task are provided in Table 9. Just as with oral production and written reception, no effect of Partner Proficiency Level was found for written productive acquisition.

Dyad type then had no effect on beginning-level learners' oral production, written reception, and written production scores. This means that the proficiency level of the Beginning-level learner's partner did not make a difference on learner's posttest performance, and that no one dyad type (Beg-Beg, Beg-Adv, Beg-NS) facilitated learning better than another. Furthermore, the possibility that gains based on dyad type (Beg-Beg, Beg-Adv, Beg-NS) are dependent on the modality in which the interaction took place (CMC vs. FTF) can be eliminated. This is because no significant effect was established for Partner Proficiency Level on beginning learners' acquisition of lexical items, regardless of modality (CMC or FTF).

Table 9. Mean scores and standard deviations (SD) for written production acquisition (depending on dyad type)

	Dyad Type	Mean	Standard Deviation	n
Written Production Pretest	Beg-Beg	.000	.000	9
	Beg-Adv	.000	.000	11
	Beg-NS	.000	.000	14
	Total	.000	.000	34
Written Production Imm. posttest	Beg-Beg	3.78	2.99	9
	Beg-Adv	4.27	2.37	11
	Beg-NS	4.21	3.29	14
	Total	4.12	2.86	34
Written Production Delayed posttest	Beg-Beg	.556	.527	9
	Beg-Adv	.636	.924	11
	Beg-NS	.714	.994	14
	Total	.647	.849	34

Discussion

In Experiment 1, Beginning-level learners of Spanish in both modalities (CMC and FTF) experienced gains in L2 vocabulary acquisition, which corroborates the results of de la Fuente (2003). Similarly, beginning-level learners in all dyad types (Beg-Beg, Beg-Adv, and Beg-NS) achieved acquisition gains. This contributes to the currently existing body of literature showing that when learners have the opportunity to negotiate for meaning and produce output, second language acquisition as measured by vocabulary is promoted. However, whether or not a participant took part in the task in the CMC or the FTF mode seemed to make a difference. Results indicate that beginning-level learners did significantly better in the CMC mode than the FTF mode on oral and written production tests, but not on receptive written production tests. This contradicts the findings of de la Fuente (2003), who found that participants in the FTF mode did better on oral and written production than those in CMC. These contrasting results may be due to the time allowed on task. In de la Fuente's study, participants were given a strict time limit in which they could negotiate the meaning of lexical items: 1 minute for the FTF group, and 2 minutes for the CMC group. Her data indicate in fact that participants seemed rushed. In the present study, participants were not limited on the amount of time they had to achieve a mutual comprehension of the items. Furthermore, in the present study, it was found that type of dyad (or proficiency level of the beginning learner's partner) did not have an effect on learning. This is relevant for pedagogical purposes when considering what types of tasks and partner's proficiency levels can be utilized to maximize acquisition.

In sum, Experiment 1 shows that negotiation for meaning can contribute to vocabulary acquisition. In addition, for beginning-level stages, computer-mediated synchronous interaction may pose more benefits than interaction in the FTF mode for developing production skills. This may be due to the fact that unlike the oral mode, computer-mediated communication is slower and therefore gives the learner more time to process and formulate (or reformulate) his or her output. Also, with whom the beginning learner is paired in a dyad does not seem to make a difference.

Performing a task in the CMC modality seems then to be highly beneficial for beginning-level learners. Upon obtaining these results, the researcher posed the question as to whether or not the saved iChat conversations from the CMC group in Experiment 1 might be used for assessment purposes. This led to Experiment 2 to explore the ways in which researchers may capitalize on some of the other valuable features of CMC. One of these features is that CMC conversations can be saved and stored for later analysis of learners' interlanguage. What might learners do with their saved L2 chat conversations?

EXPERIMENT 2

A second experiment was conducted to explore the uses of the save chat conversations. Two research questions were addressed:

1) What do learners do in going back and reviewing their saved iChat files with the researcher?

2) Can saved chat logs be used as self-diagnostic tools for assessment of L2 production?

Method

Participants

Participants for the second experiment were four beginning learners from the CMC group in Experiment 1. Two had been paired with other beginning-level learners, one with an advanced-level learner, and one with a native speaker.

Materials

The first page of each participant's saved iChat file was printed out in color to be used during the individual sessions. Participants were given pens to be able to make comments on their conversations and correct any errors they identified.

Procedure

Approximately four weeks after Experiment 1, individual sessions between the researcher and the four participants were held. During the sessions, the researcher showed the participant their saved and printed-out iChat conversation. The researcher asked the

participant to find any errors in his or her language production within the conversation. The participant was also told that he/she could highlight anything interesting in the chat log, ask questions, and discuss with the researcher parts of the conversation that may have been difficult or interesting. While most instances of highlighting errors were done by the participants, the researcher also used this time to point out some mistakes as well. The participant was encouraged to write in pen onto his/her saved iChat conversation what the correct formulation would have been. Also, participants were encouraged to ask questions that led to discussion about grammar or vocabulary. Essentially, this experiment was a session between the learner and the researcher to talk about language, identify errors in the learner's L2 output in the iChat, and assess his/her language production for pedagogical purposes. Reporting for this experiment is qualitative in nature.

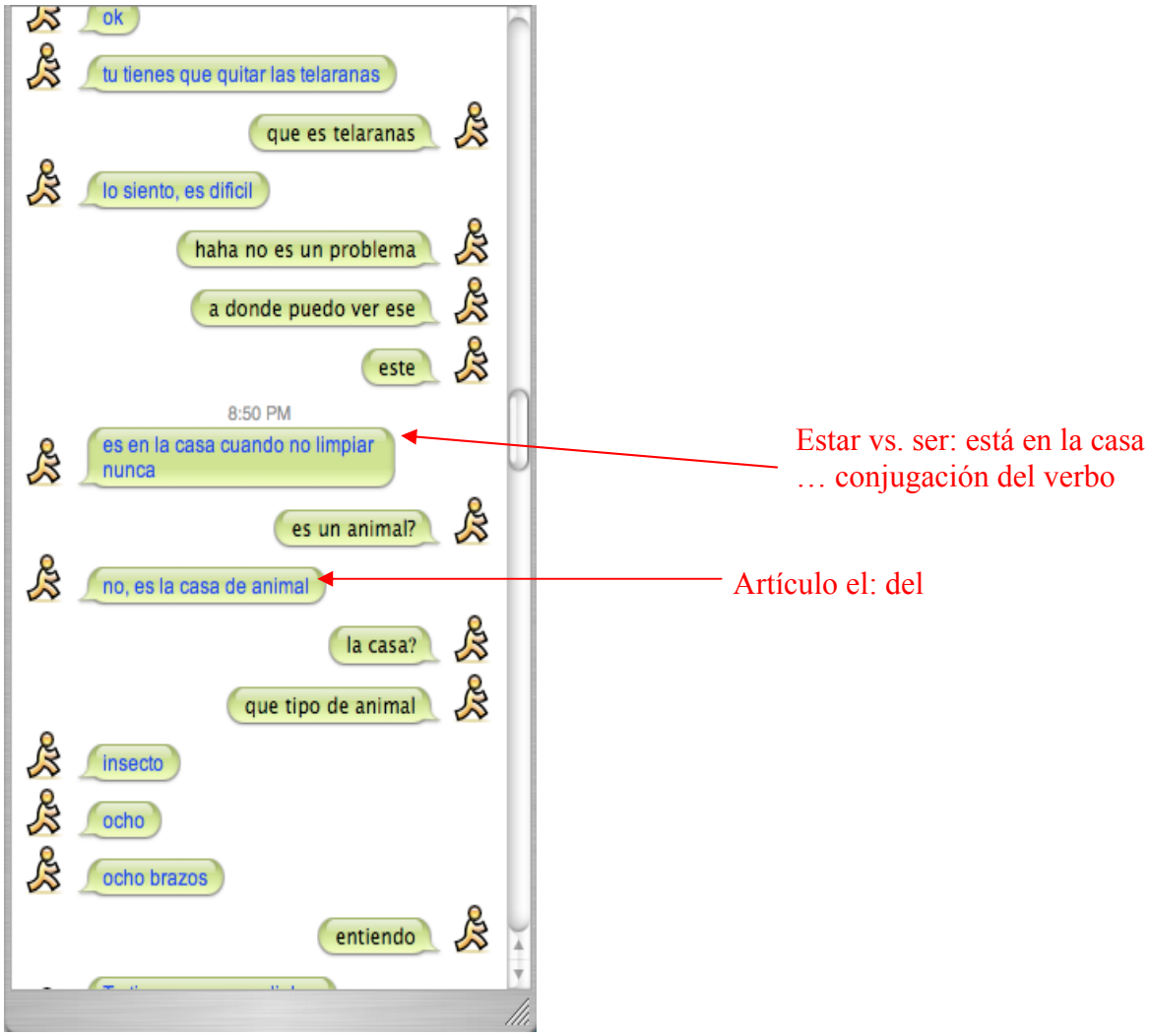


Figure 2. Learner 1 paired with Beginning Learner – self corrections

Results

Learner 1

In Figure 2, a section of the iChat conversation between Learner 1 and another beginning-level learner is provided. It shows that the beginning learner noticed some errors she had made in her output (her writing is in blue). Here, she and her partner are negotiating the meaning of the chore *quitar las telerañas* [remove the spider webs]. In one of her sentences, *Es en la casa cuando no limpiar nunca* [It is in the house when no to clean never], the learner suggested that the Spanish verb *estar* should have been used instead of *ser* (both verbs mean *to be*). Also, in the same sentence, she noticed that she had not conjugated the verb *limpiar* [to clean], leaving it in the infinitive form. In the fourth line of her speech, *No es la casa de animal* [It's not the house of the animal], the learner then pointed out that she might have used the definite article *el* [the], combining it with the Spanish preposition *de* [of] to make *del* [of the]. Notice that the corrections this learner made were not due to orthographical or accent issues, but grammatical ones. She shared with the researcher that these corrections were precisely some of the issues she had covered in class recently for composition writing. Learner 1 concluded by saying that she liked being pushed to self-evaluate her 'conversation' with the iChat file.

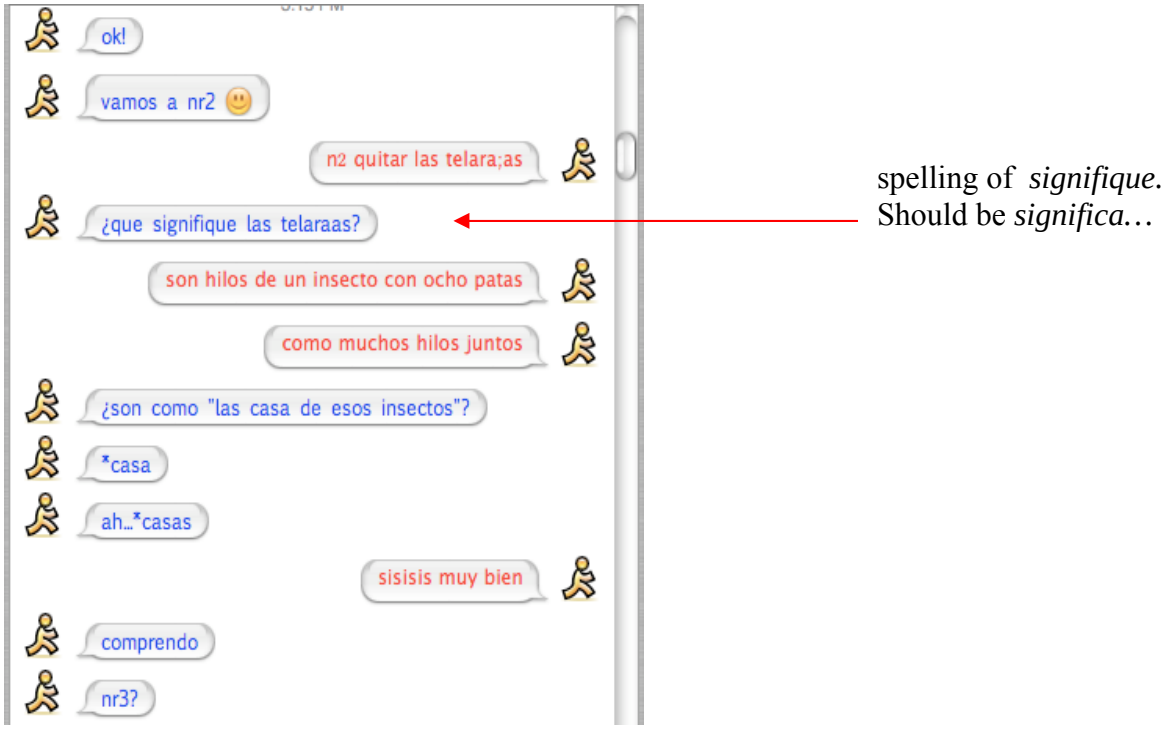


Figure 3. Recorded iChat conversation between Learner 2 and a native speaker

Learner 2

Figure 3 provides another excerpt of an iChat conversation, this time between Learner 2 (a beginning-level learner) and a native speaker. Here, the dyad is also negotiating the meaning of the chore *quitar las telerañas* [remove the spider webs].

In this section of his iChat conversation, the learner pointed out the sentence *Que signifique la teleraana?* [What does spider web mean?] He said that the spelling of the Spanish verb *signifique* [to mean] should have been *significa.* He did not highlight any other words or constructions. Two other observations were made by the researcher: notice that the learner transitions to the next chore item to be negotiated with his partner by using an emoticon ☺. This had been done earlier in the conversation by the native speaker; it might be the case that the learner was reciprocating this move and reinforcing a confirmation check. Furthermore, the learner began to use the native speaker's abbreviation for ordinal numbers, which in Spanish is done by nr1, nr2, nr3, etc. (*Nr* represent *número* [number], whereas in English this by abbreviating first with 1st, second with 2nd). Note how the learner writes *Vamos a nr2 ☺* [let's go to Number 2 ☺] in the second line and then *Nr 3?* [Number 3?] in the last line. This might be his attempt to mimic what the native speaker had been doing earlier on in the conversation (not shown). An instance of self-correction on behalf of the learner in this excerpt can also be pointed out. In the fifth line, the learner writes a sentence with *las casa.…* [the-PL house] where in Spanish, to coordinate the plurality of the article with the noun, it should be *las casas* [the houses]. He corrects himself and reformats his output in the seventh line, writing *casa, ah … *casas.* [house, ah… *houses.] He marks his reformulation with an asterisk mark in line seven. It might be the case that being able to see his written output helped the learner to notice his ungrammatical concordance with the plural form, which led to a reformulation of output.

Learner 3

In the third conversation (Figure 4), another example of self-correction done by the learner via reformulated output is provided.

In this iChat conversation excerpt, Learner 3 was paired with an advanced-level learner. In Figure 4, the dyad is negotiating the meaning of the chore *oganizar el trastero* [organize the junk room]. During the follow-up session with the researcher, this learner made specific comments and corrections about her language production in the iChat dialogue. First, as seen above, she corrects the gender of the Spanish indefinite article from *un* to *una* in the explanation "*El trastero es un parte en tu casa"* [the junk room is a part in your house]. She recognized that the article should have been feminine. The second observation she made was the lack of a relative pronoun in the fifth line, *cuarto con no usar* [room with no to use]. As can be seen, the learner had written the preposition *con,* meaning *with* or "room with no to use." She shared that she wanted to say "room that you don't use," a sentence containing a relative clause. In this same sentence, this learner also noticed that she had not initially conjugated the Spanish verb *usar* [to use]. She observed that she could have conjugated the verb for *tú* [you] (a room that you

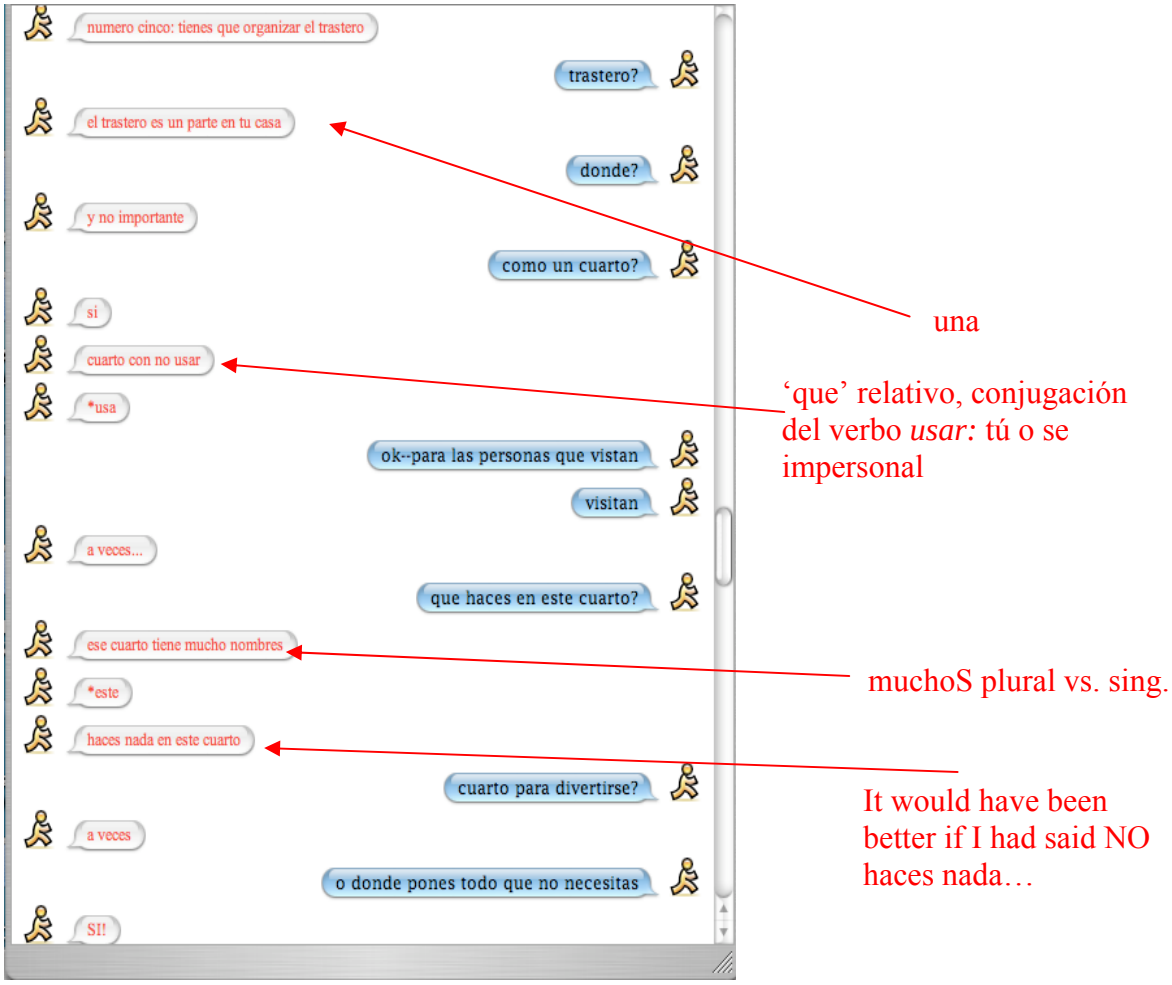


Figure 4. Learner 3 paired with Advanced Learner

don't use). Note, however, that in her original conversation, the learner did correct herself, and in the next line, conjugated the verb, writing *usa* [use]:

> *sí* [yes]
> *cuarto con no usar* [room with no to use]
> **usa* [*use] (asterisk inserted by the learner, perhaps to indicate a reformulation of output)

Learner 3 made another correction on her iChat conversation, noticing that the adjective *mucho* [many] should have been plural because it is modifying a plural noun, *nombres* [names. The last comment that this learner provided was for line ten, where she had written *haces nada en este cuarto* [you do nothing in this room]. The learner verbalized "maybe it would have been better if I had written *NO* haces nada en este cuarto [you don't do anything in this room]." In seeing her production, she recognized that *nada*

[nothing] needed the negative antecedent *no* to start the sentence, which would have made it grammatical. She and the researcher discussed this during the session and the learner made corrections on her iChat file.

Learner 4

The fourth example (seen in Figure 5) is an iChat excerpt between two beginning-level learners. In this dialogue, the learners are negotiating the meaning of the word *telerañas* (spiderwebs) from the chore *quitar las telerañas* [remove the spiderwebs]. During the session with the researcher, Learner 4 made comments and corrections on his Spanish output and also that of his partner. In line 5 of the iChat log, his partner asked *Donde son*

Figure 5. Learner 4 paired with Beginning Learner

las telerañas? [Where are the spider webs?] The learner pointed out that his partner should have used Spanish copula verb *estar* [to be] but not *ser* [to be] (to indicate physical location). The next observation that Learner 4 made was in line seven: his use of English in the iChat conversation. He had written *si o todos los places* [yes or all the *places*]. He told the researcher that at that moment, he could not think of the Spanish word for *place*. The researcher and the learner talked about this word during the session, and together they came up with other words in the target language that could have been used. The learner said he had heard and used the word before, but couldn't remember it at the time. He wrote it down as a correction on his saved iChat conversation. Learner 4 next made a comment on a sentence he had written in line 9: *Nada, estan para los piqueno animales* [nothing, they are for the small animals]. He asked if here that he had spelled the word *piqueno* [small] wrong, and wrote as a correction onto the saved iChat file *pequeño.* The learner also noticed that *pequeño* should have also been plural, as it was modifying a plural noun, and indicated this on his correction. The next correction (and question to the researcher) that the learner made referred to line 12 and 14. He had written *los animales hacen las telerannas por vivir* [the animals make the spider webs to live] and *y por tener la comida* [and for to have the food]. He asked if instead of Spanish preposition *por,* he should have used *para* in both of these instances, saying he was not sure. The learner and researcher discussed the meanings of Spanish prepositions *por* and *para* at this point; the learner took notes and wrote corrections for his errors on his saved iChat log.

Discussion

In this study, individual sessions between each of the four learners and the researcher were conducted to discuss learners' language production saved from iChat files completed during Experiment 1. Reporting of these sessions showed that CMC chat – and more importantly, being able to save iChat conversations as files – served as a unique tool for learning and reflection. The four learners discussed above were able to identify occurrences of non-understanding and errors that they had made; they were able to ask questions and notice any shortcomings in their L2 ability. Most noticing of errors noticed by the participants were grammar related, however, some questions and/or comments referenced meaning or instances of non-understanding. For the most part, participants seemed to be able to recognize problems in their own interlanguage – in one case, even that of their partner's interlanguage. Opportunities to discuss vocabulary options also arose. Overall, the sessions proved to be very beneficial, positive and insightful for both the learner and researcher. This careful examination of the chat raises the possibility that such transcripts might ultimately be used as a means of assessment for learners and also as a tool to potentially raise awareness of grammatical accuracy.

CONCLUSION

Experiment 1 showed that CMC might be a better medium than FTF to practice L2 production for the beginning-level learner, given that beginning-level learners in the CMC group had significantly higher acquisition gains in terms of vocabulary than the

FTF group. This might be because CMC does not pose the same demands for an immediate response as communication in the FTF modality does, and therefore allows for potential extra processing-time. In the CMC mode, learners can visually see language output that they and their partners produce. This reduces the cognitive demands on learners as they try to formulate language. Thus, learners have more time to think about and process what they want to say. In the mean time, learners can also test hypotheses about their L2, in that they can type out an utterance, see it, and decide whether or not it is accurate according to their interlanguage knowledge. Such hypotheses are partly confirmed by the qualitative data retrieved in Experiment 2. According to the learners, CMC allowed them to erase parts of what they wrote and reformulate their output before sending their responses to their partners. Alternatively, they sometimes noticed an ill-formed utterance and reformulated that utterance, marking it with an asterisk. For beginning-level learners, the feature that CMC is not only a synchronous chat but is also a visual stimulus might be ideal. Also, language produced by the learners in CMC can be saved and stored for analysis. This allows researchers and instructors to save and store learners' real-time language. In letting learners go back and look at their own iChat conversations, learners were able to reflect upon their performance and notice gaps in their knowledge. Whether learners spot verbs that they had not conjugated, words they did not know or had produced incorrectly, or highlight their own concordance errors, getting learners to analyze their own production in a metalinguistic way might help to further develop second language acquisition. Moreover, such transcripts might be used for pedagogical purposes by instructors who examine them to assess the stage of a learner and diagnose problems in his or her interlanguage.

REFERENCES

Abrams, Z. I. (2003). The effect of synchronous and asynchronous CMC on oral performance in German. *Modern Language Journal*, *87*(2), 157-167.

Beauvois, M. (1992). Computer-assisted classroom discussion in the foreign language classroom: conversation in slow motion. *Foreign Language Annals*, *25*(5), 455-463.

Blake, R. (2000). Computer mediated communication: A window on L2 Spanish interlanguage. *Language Learning & Technology, 4*(1), 120-136.

Böhlke, O. (2003). A comparison of student participation levels by group size and language stages during *chatroom* and face-to-face discussions in German. *CALICO Journal, 21*(1), 67-87.

Chun, D. (1994). Using computer networking to facilitate the acquisition of interactive competence. *System*, *22*(1), 17-31.

Darhower, M. (2002). Interactional features of synchronous computer-mediated communication in the intermediate L2 class: A sociocultural case study. *CALICO Journal, 19*(2), 249-277.

De la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in receptive and productive acquisition of words. *Studies in Second Language Acquisition*, *24*, 81-112.

De la Fuente, M. J. (2003). Is SLA interactionalist theory relevant to CALL? A study on the effects of computer-mediated interaction on L2 vocabulary acquisition. *Computer Assisted Language Learning*, *16*(1), 47-81.

Ellis, R., Tanaka, K. & Yamazaki, A. (1994). Classroom interaction, comprehension and the acquisition of L2 word meanings. *Language Learning*, *44*, 449-491.

Ellis, R. and He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, *21*, 285-301.

Fernández-García, M. and Martínez-Arbelaiz, A. (2002). Negotiation of meaning in nonnative speaker - nonnative speaker synchronous discussions. *CALICO Journal, 19*(2), 279-294.

Fitz, M. (2006). Discourse and participation in ESL face-to-face and written electronic conferences. *Language Learning & Technology*, *10*(1), 67-86.

Gass, S., & Varonis, E. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*, 283-302.

Iwashita, N. (2001). The effect of learner proficiency on interactional moves and modified output in nonnative-nonnative interaction in Japanese as a foreign language. *System*, *29*, 267-287.

Jepson, K. (2005). Conversations - and negotiation interaction – in text and voice chat rooms. *Language Learning and Technology*, *9*(3), 79-98.

Kern, R. (1995). Restructuring classroom interaction with networked computers: Effects of quantity and characteristics of language production. *Modern Language Journal*, *79*(4), 457-475.

Kötter, M. (2003). Negotiation of meaning and codeswitching in online tandems. *Language Learning & Technology*, *7*(2), 145-174.

Lai, C. and Zhao, Y. (2006). Noticing and text-based chat. *Language Learning & Technology*, *10*(3), 102-120.

Long, M. (1996). The role of linguistic environment in second language acquisition. In W. Ritchie and T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-438). New York: Academic Press.

Mackey, A. (1999). Input, interaction and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*, 557-587.

Meskill, C. and Anthony, N. (2005). Foreign language learning with CMC: Forms of online instructional discourse in a hybrid Russian class. *System*, *33*, 89-105.

Pellettieri, J. (2000). Negotiation in cyberspace: The role of chatting in the development of grammatical competence. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice*. Cambridge: Cambridge University Press.

Pérez, Luisa. (2003). Foreign language productivity in synchronous versus asynchronous computer-mediated communication. *CALICO Journal, 21*(1), 89-104.

Pica, T., Young, R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly*, *21*, 737-758.

Rosa, E. & Leow, R. (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *Modern Language Journal*, *88*, 192-217.

Sachs, R. & Suh, B-R. (2007). Textually enhanced recasts, learner awareness, and L2 outcomes in synchronous computer-mediated interaction. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 197-227). Oxford: Oxford University Press.

Salaberry, M. R. (2000). L2 Morphsyntactic development in text-based computer-mediated communication. *Computer Assisted Language Learning*, *13*(1), 5-27.

Sengupta, S. (2001). Exchanging ideas with peers in network-based classrooms: An aid or a pain. *Language Learning & Technology*, *5*(1), 103-134.

Shekary, M., and Tahririan, M. H. (2006). Negotiation of meaning and noticing in text-based online chat. *Modern Language Learning*, *90*(4), 557-573.

Shin, D-S. (2006). ESL students' computer-mediated communication practices: Context configuration. *Language Learning & Technology*, *10*(3), 65-84.

Smith, Bryan. (2003). Computer-mediated negotiated interaction: an expanded model. *Modern Language Journal*, *87*(1), 38-57.

Smith, B. (2004). Computer-mediated negotiated interaction and lexical acquisition. *Studies in Second Language Acquisition*, *26*, 365-398.

Smith, B. (2005). The relationship between negotiated interaction, learner uptake, and lexical acquisition in task-based computer-mediated communication. *TESOL Quarterly*, *39*(1), 33-58.

Swain, M. (1985). Communicative competence: Some roles on comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125-144). Oxford: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook on research in second language teaching and learning* (pp. 471-484). Mahwah, NJ: Lawrence Erlbaum.

Sykes, Julie. (2005). Synchronous CMC and pragmatic development: effects of oral and written chat. *CALICO Journal, 22*(3), 399-431.

Toyoda, E., and Harrison, R. (2002). Categorization of text chat communication between learners and native speakers of Japanese. *Language Learning & Technology*, *6*(1), 82-99.

Tudini, Veronica. (2003). Using native speakers in chat. *Language Learning and Technology*, *7*(3), 141-159.

Vandergriff, I. (2006). Negotiating common ground in computer-mediated versus face-to-face discussions. *Language Learning & Technology*, *10*(1), 110-138.

Warner, C. (2003). It's just a game, right? Types of play in foreign language CMC. *Language Learning & Technology*, *8*(2), 69-87.

Warschauer, M. (1996). Comparing face-to-face and electronic communication in the second language classroom. *CALICO Journal, 13*, 7-25.

Xie, Tianwei. (2002). Using internet relay chat in teaching Chinese. *CALICO Journal, 19*(3), 513-524.

[i] No specific instructions were given to participants asking that they not look up or review the target words intentionally during the period between the treatment and delayed posttest, a limitation pointed out by one of my reviewers. This is a valid point, as it refers to the problematic issue of contamination. However, a comparison of immediate posttest scores with delayed posttest scores indicates an overall decline in scores. There was no jump of accuracy, indicating that participants had probably not looked up the forms. Also, care was taken to ensure that additional exposure to the forms between test times did not take place (for example, in the normal classroom with instructors; also textbooks did not contain the forms).

[ii] The coding scheme employed in the study was dichotomous: forms were either target-like or not. Therefore, inter-rater reliability was not necessary to obtain for this experiment. My objective was to measure lexical knowledge of the target items, irrespective of minor morphological deviations.

Part IV

Authenticity in Language Assessment

Levis, J. & Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 197-208). Ames, IA: Iowa State University.

Minimal Pairs in Spoken Corpora:
Implications for Pronunciation Assessment and Teaching

John Levis
Viviana Cortes
Iowa State University

Minimal pairs, such as ship/sheep and think/sink are used in basic linguistics courses, theoretical phonology, and pronunciation teaching. One assumption underlying the use of minimal pairs is that mispronunciations in these words are likely to lead to misunderstandings. This study examines the frequencies of minimal pairs in English pronunciation teaching materials in order to examine whether miscommunication may be a result of frequency of use. Minimal pairs were collected for 26 minimal pairs of two typical pronunciation targets and related contrast sounds in ESL textbooks, /θ/ vs. /s/, /t/, /f/ (thin vs. sin, tin, fin); and /ɪ/ vs. /i/ (slip vs. sleep). These sound contrasts were used because they typify low and high functional load contrasts, according to Brown (1988). Frequencies of the words were calculated for two different corpora of spoken English: A section of the Santa Barbara Corpus of Spoken American English and the Michigan Corpus of Academic Spoken English (MICASE). Four findings are presented. First, half of the minimal pairs examined included at least one member that was extremely unlikely to occur in the corpora. Second, a surprising number of minimal pairs were potentially of the same lexical category (14/26). Third, it was unusual for minimal pairs to occur with one content word and one function word. Finally, four patterns of frequency were found in the data. We suggest that these four categories are likely to be true of all minimal pairs with other sound contrasts found in pronunciation teaching materials. Finally, four hypotheses are presented to guide future research into the effect of frequency on understanding minimal pair pronunciations.

INTRODUCTION

Minimal pairs, such as *ship/sheep* and *think/sink*, in which two words are distinguished by a single phoneme, are among the most familiar linguistic elements in basic linguistics courses, theoretical phonology, and pronunciation teaching. Minimal pairs are one of the most commonly used forms to demonstrate phonemic categories in any language, and have therefore played an important role in for linguists as they establish the meaningful elements of language. Not only have they been theoretically useful, they are a mainstay for teaching pronunciation through their use in pronunciation diagnostic assessment, spoken language production practice, and listening comprehension materials. Brown (1995) notes that exercises using minimal pairs are ubiquitous in pronunciation teaching materials. Minimal pairs seem an obvious choice for diagnostic assessments aimed at

identifying specific areas of language knowledge for learners to work on. Our own informal survey of current American pronunciation textbooks confirms that minimal pairs play an important role in all but one.

An assumption driving the use of minimal pairs in teaching is that foreign language learners, by using the wrong sound, are more likely to be misunderstood because listeners will be led to believe that another word was intended. Berlitz Language Schools plays on this assumption in a video advertisement for English classes. The ad shows a young officer in a dimly lit room surrounded by glowing equipment. An older officer, speaking German, gives the young man instructions about the job he will do, then leaves the room. As the young man waits at the console, a distress call comes over the radio (http://www.youtube.com/watch?v=f9cv0dRLsUM).

Voice:	Mayday, Mayday! Hello. Can you hear us? Can--you--hear us? Can you [static] ? Over. We are sinking. We are---sink--!
Young man:	Hallo, zis is ze German Coastguard.
Voice:	We're sinking! We're sinking!
Young man:	What are you sinking [thinking] about?

This advertisement sends three messages:

- The inability to distinguish two phonemes leads to a loss of intelligibility.
- Loss of intelligibility has rather serious repercussions.
- A particular sound, in this case /θ/, is a serious problem.

The ad cleverly uses the one possible linguistic context in which these two sounds could create such a misunderstanding, but how frequently do such linguistic contexts appear in normal spoken language? Even though it is possible to create a scenario in which the consequence for mispronunciations could be horrendous, such contexts are not necessarily common. The utility of minimal pairs for advertisers developing clever ads and for linguists determining phonemic distinctions does not necessarily imply their utility for language teaching and assessment. Such language education functions need to be considered in part on the basis of the communicative problems that are likely to result for learners who fail to make appropriate phonemic distinctions in the words appearing in minimal pairs.

This study explores the utility of minimal pairs in language learning and assessment materials by examining the use of the words contained in English minimal pairs in two corpora of spoken American English. For example, the pair "should-shoed" appears in text books, but if the occurrence of "shoed" in spoken language is as infrequent as one might suspect, there is little, if any, possibility that a mispronunciation of "should" in a sentence such as "We should get to the airport by 6:00" will result in miscommunication due to the listeners' confusion of "should" with "shoed." This example reveals another reason that such a miscommunication implausible: The syntactic place and role of the word "should" provides good cues to the listeners that help them to interpret the message despite an error in the vowel pronunciation. The plausibility of miscommunication

occurring due to mispronunciation of a key sound contrast needs to be assessed through the examination of frequencies and grammatical categories of actual minimal pairs used in learning materials. The purpose of this study is to do just that. On the basis of these findings, we provide hypotheses for further research into the importance of minimal pairs for intelligibility.

MINIMAL PAIRS IN PRONUNCIATION TEACHING

Minimal pairs are the backbone of the teaching of vowel and consonant sounds in ESL pronunciation texts. They are featured in both listening and production exercises. In listening exercises, learners hear one or both members of minimal pairs that are particularly hard for them to distinguish. For example, exercises may ask learners to identify whether two words are the same or different, which word of the two contains a particular sound (e.g., Which has the /i/ sound, the first word or the second (*Seat* or *Sit*), which word of three is different from the other two, and whether a word has a particular targeted sound. In these types of exercises, very few restrictions are evident on the words used. Pairs do not have to be of the same lexical category, they do not need a context, and the words do not need to be common. For example, in the pair, *thigh/thy*, sometimes used to illustrate the difference in the two English "th" sounds, the words are of different categories. It is difficult to find a context in which both are likely. Both words are also uncommon, and *thy* is restricted to contexts in which Middle English forms are likely to appear (such as some religious services or reading Shakespeare).

Slightly more meaningful uses of minimal pair exercises are found in books like Grant (2001, p. 194), reproduced in Example 1. This exercise asks one student to read one of the two prompts while the other student responds appropriately. An inappropriate response results in a communication breakdown, which prompts both students to be more careful about their pronunciation and their listening.

Example 1

	PROMPTS (STUDENT 1)	RESPONSES (STUDENT 2)
a.	Did you slip?	(Yes, on the ice.)
	Did you sleep?	(Yes, for 10 hours.)
b.	Those were beautiful pitches.	(It was a great baseball game.)
	Those were beautiful peaches.	(It was a good crop.)

Unlike the non-contextualized minimal pair exercises, this kind of exercise requires minimal pairs which are both the same part of speech (verbs in the first example, nouns in the second), equally likely in the same linguistic context, and semantically plausible, a requirement which "is not possible for the majority of minimal pairs in English (Brown, 1988, p. 601).

PREVIOUS RESEARCH

Even when minimal pairs are members of the same lexical category (e.g., *ship* and *sheep* are both nouns), they are rarely equally likely in the same context (Brown 1995). Moreover, the context or likely collocations often makes one word far more likely in a listener's interpretation (Cruz, 2005), minimizing the possibility of misunderstanding. Jenkins' research (2000), however, suggests that top-down processing effects implied by things like knowledge of collocations are more likely for native than nonnative listeners.

Measuring the relative importance of minimal pairs is difficult. In one of the only useful treatments of this problem, Brown (1988) provides a modified measure of frequency in describing the functional load (FL) of various minimal pairs that occur in textbooks for teaching pronunciation. Functional load "is a measure of the work two phonemes do in keeping utterances apart" (King, 1967, as cited in Munro & Derwing 2006, 522). Brown calculates a ranking that takes into account 12 factors (such as the number of initial minimal pairs that exist for a given contrast, the number of final minimal pairs, and the likelihood that the distinction is enforced in all varieties of English). Brown says that "perhaps the most difficult [issue] to find a satisfactory solution to is that of the relative weighting of the12 factors" [that can be used to modify raw frequency counts] (p. 603). Brown, however, does not give us the details of the weighting he used for each factor. Several of Brown's factors that are most relevant to this study are as follows.

1. What is the relative probability of the sounds occurring? (e.g., /ɪ/ is four times as likely to occur as /i/)
2. How many minimal pairs exist for a contrast? (e.g., few possible pairs exist for /ʊ/-/u/)
3. How frequent are the members of a minimal pair? As mentioned, few possible pairs exist for /ʊ/-/u/. Those that do exist have such uncommon words for /u/ (e.g., shoed, wooed, cooed, Luke) that they may as well not exist. One author states that "the functional load of a contrast depends on the existence of minimal pairs of words that are both frequent" (quoted in Brown, 1988, p. 601).
4. How many minimal pairs for a sound contrast belong to the same part of speech?
5. How many minimal pairs can occur in the same semantic context?

When Brown wrote, all of these questions could be answered either through phoneme frequency counts or through intuition. In contrast, no studies have examined the frequency of minimal pairs used in pronunciation teaching materials through the use of spoken corpora. Thus the goal of this exploratory study is to examine whether the minimal pairs used to teach common problem sounds are frequent in actual usage.

Munro & Derwing (2006) tested Brown's FL predictions using two low FL pairs (/θ/-/f/;

/ð/-/d/) and two high FL pairs (/l/-/n/; /s/-/ʃ/). In the study, native English speaking listeners heard eight types of sentences: those with 0, 1, 2, and 3 errors in words with low FL; 0, 1, and 2 errors in words with high FL; and 1 low FL error and 1 high FL error in the same sentence. Sentences were judged for accentedness and comprehensibility, both using a 9 point scale. Both sets of judgments showed that listeners reacted much more strongly to high FL errors than to low FL errors.

In the accentedness ratings, sentences with 1, 2, and 3 low FL errors were heard as more accented than sentences with no errors, but there was no cumulative effect of frequency for low FL errors. Thus, any type of error increases accentedness judgments but accentedness does not increase with greater error quantity. In addition, sentences with 1 or more high FL errors were always heard as more accented than sentences with any number of low FL errors. For high FL errors, there was a cumulative effect of frequency, that is, sentences with 2 high FL errors were heard as more accented than sentences with 1 high FL error.

Comprehensibility ratings, which measure listeners' perception of how easy a speaker is to understand, found similar results. Sentences with 1, 2, and 3 low FL errors were perceived as less comprehensible than those with no errors. Again, there was no cumulative effect of frequency for low FL errors. Sentences with high FL errors were always perceived to be less comprehensible than sentences with any number of low FL errors. However, there was no cumulative effect of frequency for high FL errors. Sentences with 1 or 2 errors were rated as equally (in)comprehensible.

METHOD

For this study, minimal pairs used in common pronunciation texts were collected (Grant, 2001; Dauer, 1993; Orion, 1997; Lane, 1993; Miller, 2000). We concentrated on two typical pronunciation targets and related contrast sounds in ESL textbooks, /θ/ vs. /s/, /t/, /f/ (thin vs. sin, tin, fin); and /ɪ/ vs. /i/ (slip vs. sleep). The /θ/ pairs are examples of low FL errors according to Brown (1988), with the following functional loads (on a 10-point scale): /θ/-/f/ =1; /θ/-/t/ =4; /θ/-/s/ =5. The calculations Brown used are not given in his article. The second set is an example of a high FL error, /ɪ/ vs. /i/ (slip vs. sleep), with an FL =8. There were 26 minimal pairs identified from the textbooks for these sounds, 16 for /θ/ vs. /s/, /t/, /f/ and 10 for/ɪ/ vs. /i/ Grant, 2001; Dauer, 1993; Orion, 1997; Lane, 1993; Miller, 2000).

In order to check the frequency of these words in natural spoken language, we investigated two different corpora of spoken English. One was a section of the Santa Barbara Corpus of Spoken American English (Du Bois, Chafe, Wallace, Meyer, & Thompson, 2000) which we obtained through the Iowa State University Library's subscription to the Linguistic Data Consortium (LCD). This corpus is based on a large body of naturally occurring spoken interactions recorded all over the United States. It includes language produced by speakers with different regional origins, and of different ages, occupations, genders and ethnic and social backgrounds. Face-to-face interactions

are the predominant form of language represented but the corpus also presents telephone conversations, sermons, story-telling and other forms of spoken language. The other corpus was the Michigan Corpus of Academic Spoken English (MICASE), a collection of about 200 hours of academic speech that was recorded and transcribed at the University of Michigan (Simpson-Vlach & Leicher, 2006). This corpus is made up of language recorded from in- and out-of class events such as lectures, student presentations, office hours, and service encounters as well as many other speech events frequently encountered in university life. In the case of the Santa Barbara Corpus, a computer program included with the corpus identified the target words in the minimal pairs, counted their occurrences in the corpus, and normalized the frequencies to 100,000. To identify frequencies in MICASE, the online concordancer was used to identify frequencies and then normalization was done in an Excel database. All frequencies from both corpora, both raw and normalized, were transferred to an Excel spreadsheet. Appendix A shows the frequencies corresponding to each corpus. In addition to these corpora, other supplementary corpora were analyzed to identify the frequency of these minimal pairs (a corpus of university lectures and a corpus of everyday conversation). These searches yielded the same relationship between the relative frequencies of the members of the pair/group of words investigated.

RESULTS

Four main descriptive findings are notable. First, half of the minimal pairs examined had at least one member that was extremely unlikely to occur in spoken corpora. That is, 13 of the 26 pairs examined included a member that was rare, and so very unlikely to be familiar to learners of English. This suggests that many minimal pairs in the textbooks probably fail a very basic test of usefulness. Second, a surprising number of minimal pairs were potentially of the same lexical category (14/26). Thus it seems that it may not be overly difficult to find minimal pairs that overlap in this way. Third, it was unusual for this selection of words to have minimal pairs with one content word and one function word. Because function words and content words do not play the same grammatical roles in sentences and because they have differing rhythmic patterns in spoken discourse, it seems they are less likely to be a source of confusion. Finally, four patterns of frequency were found in the data. We suggest that these four categories are likely to be true of all minimal pairs with other sound contrasts found in pronunciation teaching materials. The sound contrasts chosen for this pilot study are used in all textbooks, and represent sounds that are found across the functional load scales. We believe it would be surprising if other minimal pair contrasts in teaching materials did not follow similar patterns, though this possibility must be left open until a fuller sampling is analyzed.

The first category, which we call Group A, included one member which was very common and one which was very uncommon. Eight minimal pairs of the 26 fit this description, as shown in Table 1, illustrated by *think* and *sink*. *Think* is extremely common in these four corpora, while *sink* is almost nonexistent. Assuming that listeners are most likely to interpret a word in part because of its frequency, this calls into question

the likelihood of a misunderstanding such as the one portrayed in the Berlitz commercial.

The second category included one member which was very common and one which was less common, but not extremely uncommon as in Group A. Seven of 26 pairs fit this category, as illustrated by the minimal triple in Table 2. These minimal pairs may be more likely to cause misunderstanding because they are all somewhat common.

Table 1: Group A pairs (very common and very uncommon)

	SBC	MICASE
think	133	6188
sink	0	7

Table 2. Group B pairs (very common and somewhat common)

	SBC	MICASE
three	68	1664
tree	14	411
free	13	166

Table 3. Group C pairs (equally common)

	SBC	MICASE
leave	12	149
live	14	130

Table 4. Group D pairs (equally uncommon)

	SBC	MICASE
peel	0	9
pill	1	6

Category C pairs included pairs which were equally common. Six pairs were included in this category, as shown in Table 3. These pairs included members neither of which was very common but which were not rare. In other words, both members were fairly frequent words with roughly equivalent chances of occurring in the corpora.

Category D pairs were equally rare. Five pairs, illustrated by *peel/pill*, were included in this category (Table 4). These pairs almost never occurred in the corpora, indicating that they had very little likelihood of occurring in normal spoken language, and thus are probably not good candidates for teaching.

DISCUSSION

This section will look first at controversies surrounding the teaching of /θ/, then will examine the results for /ɪ/ vs. /i/, discuss some implications for pedagogy, and will provide four hypotheses that can guide further research into importance of minimal pairs for intelligibility.

The sound /θ/ has been the target of much argument in pronunciation teaching. A distinctively English sound which is shared with few other languages, /θ/ is one of the most commonly taught sounds and the one sound which most ESL learners feel they should learn to pronounce. In contrast, many theorists argue that /θ/ should not be taught, a recommendation that fits with Brown (1988). Jenkins (2000) says that /θ/ should not be taught as it rarely caused misunderstanding in her study of NNS-NNS interactions. Another reason that /θ/ should not be taught is that native speaker varieties often use variant pronunciations, especially the /t/ and /f/.

In this study, we found that /θ/ words are usually very frequent but their minimal pair is not (e.g., think/sink; thank/sank; through/true). The infrequent misunderstanding of /θ/ words may be because of the unlikely occurrence of the minimal pair with which it might be confused.

This hypothesis gains some support in light of findings from Deterding (2005). He says that /θ/ may be important for listening to native speech since certain variants can confuse nonnative listeners. Deterding studied the ability of NNS listeners in Singapore to understand Estuary English speakers who regularly used /f/ rather than /θ/. The NNS listeners, who expected /θ/, found it difficult to interpret words like “three” because of their minimal pair “free.” As shown earlier, this particular pair includes two words that are both somewhat common in the corpora.

The other minimal pair studied, /ɪ/ vs. /i/, showed diametrically opposed patterns of frequency. In 5/10 pairs, both members were relatively common, and in 5/10 pairs, one or both members were very uncommon. Even though this minimal pair contrast is considered high FL, it is unlikely that these two patterns will result in a similar likelihood of misunderstanding because we believe that frequency of occurrence is an important factor in whether misunderstandings are likely.

What can be learned from these frequency counts? First, minimal pairs are used for three different purposes in pronunciation teaching: to determine learners' ability to hear contrasts, for listening comprehension practice, and for spoken language production. Minimal pair exercises may be a useful way to determine whether learners can hear particular sound contrasts, as these items provide a quick way to determine whether learners can hear differences between sounds in the target language. It should also be clear that pairs that include relatively frequent items should be used in this way, especially if they can be put into a context in which both members of the pair can occur. Minimal pairs may also be useful for micro-level listening practice, especially if the listening is contextualized and used with relatively common words while avoiding pairs in which one or both members are rare. This means that only a small number of those minimal pairs currently in pronunciation books should be used in pronunciation teaching. It is less clear as to whether minimal pair exercises are helpful for speaking. Misunderstandings in natural speech are rarely a result of minimal pairs with no other factors. There is little evidence for the assumption that the mispronunciation of one sound will be enough to irretrievably harm understanding.

HYPOTHESES FOR FUTURE RESEARCH

Four hypotheses can be derived from this study and used to guide future research into the effect of frequency on using minimal pairs. The first two hypotheses relate to pairs that are unlikely to lead to misunderstanding, and the last two to pairs that we believe are more likely to lead to difficulties. First, we suggest that pairs that include one content word and one function word (e.g., eat/it) are unlikely to cause problems for listeners, regardless or frequency, as the function of the words in sentences is very different. Second, we suggest that if one word in a pair is extremely likely and the other is extremely unlikely, misunderstandings are unlikely to occur, regardless of word class.

Third, if both words in a pair are relatively (un)likely, listeners will be more likely to misinterpret. This may not lead to misunderstanding if both members of a minimal pair are rare in spoken language because of the lack of likelihood that the words will occur.

Fourth, the greatest likelihood of misinterpretation will come when both words are of the same lexical category, are relatively frequent, and are semantically plausible. This is fairly unusual, though by no means impossible. It may even be that pairs which are not of the same lexical category can still cause problems if the members of the pair are relatively frequent and semantically plausible. This is because our arguments are based on the assumption that learners' minimal pair errors are the only errors they make, and that errors in syntax, morphology, pragmatics, etc. do not enter into the equation. This is obviously not the case in most spoken language produced by NNSs. The listener has a much bigger job than simply decoding the difference between two sounds. Enough mistakes in an utterance, regardless of whether minimal pairs are involved, can make even the most tolerant listeners send out a cry of "Mayday!" when they think their understanding is sinking.

REFERENCES

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22*(4), 593-606.

Brown, A. (1995). Minimal pairs: minimal importance? *ELT Journal, 49*(2), 169-175

Cruz, N. (2005). Minimal pairs: Are they suitable to illustrate meaning confusion derived from mispronunciation in Brazilian learners' English? *Linguagem & Ensino*, *8*(2), 171-180.

Dauer, R. (1993). *Accurate English*. Englewood Cliffs, NJ: Regents Prentice Hall.

Deterding, D. (2005). Listening to Estuary English in Singapore. *TESOL Quarterly, 39*(3), 425-440.

Du Bois, J., Chafe, W., Meyer, C., & Thompson, S. (2000). *Santa Barbara corpus of spoken American English, Part 1.* Philadelphia: Linguistic Data Consortium.

Grant, L. (2001). *Well said* (2nd ed.). Boston: Heinle & Heinle.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

King, R.D. (1967). Functional load and sound change. *Language, 43*, 831-852.

Lane, L. (1993). *Focus on pronunciation.* New York: Longman.

Miller, S. (2000). *Targeting pronunciation.* Houghton-Mifflin.

Munro, M., & Derwing, T. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520-531.

Orion, G. (1997). *Pronouncing American English* (2nd ed.). Boston: Heinle & Heinle.

Simpson-Vlach, R., & Leicher, S. (2006). *The MICASE handbook.* Ann Arbor: The University of Michigan Press.

Appendix A

	Santa Barbara corpus (approx 140,000 words)		MICASE (approx. 1,800,000 words)	
	Santa Barbara (raw counts)	Santa Barbara (normed to 1,000,000)	MICASE (raw counts)	MICASE (normed to 1,000,000)
think	133	950	6188	3437
sink	0	0	7	4
thick	3	21	39	22
sick	3	21	40	22
tick	5	36	6	3
thought	44	314	971	539
sought	0		7	4
taught	3	21	64	35
fought	0	0	6	3
thank	25	178	460	255
sank	0	0	1	0.5
tank	0	0	38	21
thin	2	14	52	29
sin	0		13	7
tin	1	7	4	2
fin	0	0	44	24
three	68	486	1664	924
tree	14	100	411	228
free	13	93	166	92
mouth	5	36	53	29
mouse	3	21	48	27
worth	14	100	117	65
worse	1	7	77	42
through	48	343	829	460
.true	8	57	470	261
each	9	64	829	460
itch	0	0	3	2

eat	22	157	124	69
it	1000	7143	35105	19507
deed	1	7	0	
did	185	1321	226	125
sleep	5	36	23	13
slip	0	0	22	12
				0
feet	10	71	65	36
fit	4	28	147	82
leave	12	86	268	149
live	14	100	234	130
peach	0	0	2	1
pitch	0	0	9	5
peel	0	0	9	5
pill	1	7	6	3
heat	11	78	109	61
hit	12	86	111	62
seen	13	93	369	205
sin	1	7	13	7

Pardo-Ballester, C. (2008). The construct validity of a web-based listening comprehension exam. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 209-227). Ames, IA: Iowa State University.

The Construct Validity of a Web-Based Listening Comprehension Exam

Cristina Pardo-Ballester
Iowa State University

It is well known by test developers that the construct to be measured by a test needs to be clearly defined as part of the process of developing a test (Bachman, 1990; Davidson and Lynch, 2002). This study focuses on a trial version of an online Spanish Listening Exam (SLE), a listening measure focused on grammatical items and tasks based on the main topics learned in the first two years of a Spanish curriculum. The SLE tasks are relevant to the language instruction domain which helps to define the test construct. This paper describes research evaluating the construct validity of the SLE through research on content relevance, criterion-relatedness and content coverage.

INTRODUCTION

The Spanish Listening Exam (SLE) presented in this paper is an exam based on tasks that are typically used for instruction in Spanish language classes, including normal features of spontaneous spoken discourse such as false starts or hesitations. The SLE is a linear exam in which Item Response Theory (IRT) was used in order to improve the test reliability and validity of the test. Several types of data were used as evidence of construct validity. Buck's (2001) theory-based definition of language knowledge was used to interpret the scores in terms of listening ability. Content coverage evidence is demonstrated through item and task congruency with the instructional content of our elementary and intermediate language program. Also, content relevance evidence is obtained by consulting content experts' opinions.

The purpose of the SLE is to group learners according to their levels of language ability for learning Spanish. Three final cut scores were used to place students into three different proficiency level groups. To set these cut scores two standard-setting procedures were implemented: the bookmark method and the borderline-group method.

CONSTRUCT OF LISTENING

The construct of the SLE test is based on an interaction between the listening ability, the tasks, and the Spanish course syllabi used at the University of California, Davis (UCD). Since syllabi are altered over the years to reflect changes in textbooks, thirteen Spanish

textbooks were examined in order to identify the main topics (grammar and vocabulary) that normally are taught in the first-and second-year of the Spanish language curriculum. The *Dos Mundos* (Terrell, Andrade, Egasse, & Muñoz, 1998) and *Al corriente* (Blake, González Pagani, Ramos & Marks, 2003) textbooks are used in UCD elementary and intermediate language program, and they were used as the main sources of information. However, ten more textbooks were also used as relevant teaching materials (see Pardo Ballester, 2007). To measure the Spanish L2 listening ability, this study examines two components of Buck's (2001: p. 104) framework which were included in the SLE:

> Grammatical knowledge: the ability to understand short utterances on a literal semantic level. This includes phonological modification, spoken vocabulary, and spoken syntax, expressive intonation, and stress. Items in the SLE are categorized according to item difficulty—local or inference—to test the most salient phonological, lexical and syntactic features presented in introductory and intermediate Spanish textbooks as a foreign/second language. Shohamy and Inbar (1991) showed that oral texts identified as informal spontaneous speech are easier to understand than those texts identified as formal written speech. In their study they identified local questions (i.e., understanding single words, facts or locate details from the passage), trivial items (i.e., recalling details as names or numbers from memory) and global questions (i.e., drawing conclusions). They found that participants who answered the global questions correctly also received correct answers for the local items. In general, global items were more difficult to answer appropriately than local items. On the other hand, trivial items showed mixed results and Shohamy and Inbar suggested not using this type of items in listening comprehension tests. According to Tsui and Fullilove (1998), bottom-up processing is fundamental to discriminate among the listening performance of L2 learners. Less-skilled L2 listeners are weak in bottom-up processing because they lack automatized linguistic decoding skills.

Sociolinguistic knowledge: understanding the language of particular sociocultural settings. The difficulty of the oral texts is measured according to idiomatic expressions and dialectal and cultural references.

Based on the SLE construct three hypotheses were stated:

1) Students placed in higher levels would score better than those placed in lower levels.

2) Items coded as comprehension would be more difficult than local items. Lexical items would differ from the phonological and syntactic items.

3) Tasks classified as the most difficult based on the sociolinguistic features would demonstrate higher levels of difficulty on the SLE.

TEST TASKS

Because the test's purpose is to place students in the Spanish lower-division program, the SLE includes a variety of tasks based on semi-scripted oral text types and local or inference items. With the oral stimuli, ten different tasks for the Spanish listening exam were built with a range of six to fourteen items per task. All items were scored from 0 to

1 on grammatical knowledge. Items were true/false, multiple-choice items and limited production questions. The limited production questions call for simple and minimal responses, so that the test takers' writing skills would affect performance only minimally.

Test-takers preview the items before listening to the oral input. Items are intended to be easy or difficult depending on the learners' levels of proficiency. For example, after having listened to a passage about a cultural tradition, participants hear an unfamiliar word *lustrosos* 'shiny' and they read *brillantes* 'shiny' when they were asked a true/false question which is reproduced below:[i]

Item 77: Local lexical

Los zapatos están muy brillantes	*Cierto*	*Falso*
'The shoes are very shiny'	True	False

This lexical item designed for second year can be difficult for a first-year learner to answer correctly, because it is considered part of low-frequency vocabulary. By the same token, it is expected that most learners will be able to succeed with an easy item (i.e., high frequency vocabulary) designed for first-year learners (e.g., they hear *amiga 'friend'* and they read *novia 'girlfriend'*, two familiar words for all levels of learners). As an example of a phonological item, consider when students hear a word they do not know, such as *bachillerato* 'high school diploma' they may select an answer that sounds similar and is part of their lexical knowledge such as *barato* 'cheap' as reproduced here:

Item 39: Local phonological

¿Qué hace su amigo?	What does his friend do?
a. *Trabaja en una tienda*	a. He works in a store
b. *Trabaja para el gobierno*	b. He works for the government
c. *Estudia bachiller* diploma	c. He studies to receive his high school
d. Estudia muy barato	d. He studies at a very cheap rate

A phonological item requires the ability to also produce factual answers in the form of precise names, or numerical details from memory, such as *diecinueve* 'nineteen' or *veintinueve* 'twenty-nine'.

The speed and regional dialect of the speakers during different text types is also considered to contribute to the item difficulty and therefore, beginning learners could only answer a few items correctly based on familiarity with the topic (e.g., *La fiesta de los Reyes Magos* 'an Epiphany holiday'). Buck (2001) mentions that, in order to respond to items, learners should be dependent on listening to the oral passage. However, the SLE is a multi-level test and some tasks are more difficult for beginners due to the speech rate and pronunciation. Therefore, a decision was made to include items that allow students to use their background knowledge to respond, in order to avoid frustration for beginners. The number of speakers is also taken into consideration when classifying oral texts based on difficulty level.

Based on the intended content coverage of the test, three content-relevance hypotheses were investigated.

1) Raters will find the content of the test to correspond to the content of the course book.

2) Raters will find the difficulty of the test tasks to correspond to the appropriate ACTFL level.

3) Raters will find the overall characteristics of the test tasks to be appropriate for the level of the student who will take the test.

METHOD

Participants

The results of this study are intended to generalize to potential learners in the lower level division courses. The sample consisted of 147 students enrolled in different Spanish classes at the University of California, Davis. The breakdown of the numbers of participants at those courses is shown in Table 1 corresponding to the ACTFL proficiency levels.

Materials

WebLAS is the acronym of the Web-based Language Assessment System. WebLAS was constructed by programmers working with test developers at UCLA (UCLA Department of Applied Linguistics and TESL & Center for Digital Humanities, 2003). A collaborative project carried out at UCLA resulted in the design and development of placement exams in ESL, Korean and Japanese including listening, writing and reading. The development of the Spanish listening exam was part of the WebLAS project carried out at UC Davis. Lyle Bachman, the principal investigator of WebLAS, allowed us to use WebLAS for research. With our collaboration with the development of the SLE, UC Davis contributed with feedback about possible problems encountered during the use of WebLAS.

Table 1. Proficiency level students for the SLE

Proficiency Level	Number of students	Percentage
Novice-high	34	23%
Intermediate-low	60	41%
Intermediate-mid	53	36%

Procedures

Eleven Spanish native speakers from Argentina, Mexico, Peru, and Spain were recruited for the production of the recorded materials. Having a variety of different speakers with different accents was important because of the diversity of Spanish teachers at UC Davis.

Items were created from listening to the oral input rather than reading the text because this process ensured creating items which were focused on comprehension rather than difficult items which are normally based on memory of small details. To trial these new tasks I visited various classrooms to administer the listening tasks with paper and pencil. These tasks were considered to be a listening practice for students. Once they answered all items they were asked to circle the words that they had trouble understanding and then add comments on the difficulty of the tasks. The Spanish instructors from those classes were also asked to collaborate by answering the items and adding comments concerning the difficulty of items according to their perception and own classroom experience. The rationale for asking students and instructors about the test difficulty was to provide support for the exam's content validity.

The passages for measuring listening comprehension were ranked with the ACTFL (1986) guidelines in mind. The difficulty of passages was ranked according to the following parameters: 1) grammar points, 2) suggestions given by the ACTFL proficiency guidelines, 3) concrete or abstract content, 4) rate of delivery, 5) number of people speaking, and 6) idiomatic expressions. To contribute to the content validity of the tasks, four graduate students were asked to evaluate the oral stimuli. The evaluator materials contained an evaluation form with three different sections (see Appendix A).

The SLE was administered and taken in the computer language laboratory. Access to the Internet Explorer program was working properly on the twenty-five computers. All the headphones were functional. The instructions for the online test were administered. Test takers were informed that they had 40 minutes to complete the test, but they could complete it in less time and that guessing would not affect their scores. Their task was to interact in Spanish with WebLAS, completing ten listening tasks by listening to oral stimuli, reading the context and items and then selecting or producing the right answer.

Analysis

After pretesting the items, the Rasch model was used to calibrate the items using WINSTEPS (Linacre, 2006). Data were also analyzed using the SPSS 12.00 (2003) package for analyzing and interpreting the relationship of the ability and difficulty measures to other factors such as task level, item type, and year of proficiency. The construct validity of the SLE is investigated by examining the interactions among the dependent and independent variables. Univariate one-way ANOVA was used to investigate the differences in difficulty of items and task levels. The analyses also compared the difficulty of proficiency levels and students performance using IRT estimates from WINSTEPS.

RESULTS

The content coverage and criterion-relatedness of the SLE is evaluated by different analyses of variance. Descriptive statistics show how the content relevance of the SLE is examined by using expert judgments. The purpose of these analyses is to provide evidence for the validity of the intended interpretations from the SLA scores.

Content

One way to address the content aspect of construct validity is by asking experts to appraise the assessment tasks. In this study four instructors were asked to rate the ten listening tasks of the SLE. The evaluation was done in three phases following our three content-related hypotheses about the correspondence of the items to the course book, their reflection of the ACTFL levels, and the level appropriateness for the intended test takers. (See Appendix A).

Descriptive statistics were calculated for all rating scale data in order to determine score distribution patterns of the content relevance according to the raters' evaluation. Table 2 shows descriptive statistics for each rater using the text content scales (1=strongly disagree to 5=strongly agree) to judge the agreement of test content for each task belonging to one group (See Appendix A, section A) based on the linguistic features included in each text which were found in Spanish textbooks. The software SPSS was not able to calculate kurtosis and skewness statistics for rater 2 because the mean was 5.00 and the standard deviation was 0.00 indicating no variation. The kurtosis and skew use the standard deviation in the denominator, so they are not defined because variation does not exist for this rater.

Table 3 shows descriptive statistics for each rater using the content scale of 1=Novice-high, 2=Intermediate-low, and 3=Intermediate-mid to judge the agreement of oral passages for each task belonging to one group (See Appendix A, section B) based on the ACTFL proficiency levels. This general pattern seems to indicate a fair amount of consistency among raters in applying the ACTFL guidelines to the level of difficulty/ability.

Table 2. Descriptive statistics for raters on content ratings based on Spanish textbooks (scale 1 to 5)

Raters	N	Mean	SD	Min	Max	Skew	Kurtosis
Rater 1	10	4.50	.707	3	5	-1.179	.571
Rater 2	10	5.00	.000	5	5	.	.
Rater 3	10	4.90	.316	4	5	-3.162	10.000
Rater 4	10	4.90	.316	4	5	-3.162	10.000

Table 3. Descriptive statistics for each rater on oral passages based on the ACTFL guidelines (scale 1=Novice-high to 3=Intermediate-mid)

Raters	N	Mean	SD	Min	Max	Skew	Kurtosis
Rater 1	10	1.90	.737	1	3	.166	-.734
Rater 2	10	1.90	.737	1	3	.166	-.734
Rater 3	10	1.70	.676	1	3	.433	-.283
Rater 4	10	1.90	.567	1	3	-.091	1.498

Table 4. Descriptive statistics for each rater on oral passages based of rate speech, textbooks, and ACTFL guidelines and other attributes. (Scale 1=easiest text to 3=most difficult text)

Raters	N	Mean	SD	Min	Max	Skew	Kurtosis
Rater 1	10	1.90	.738	1	3	.166	-.734
Rater 2	10	1.90	.738	1	3	.166	-.734
Rater 3	10	2.00	.816	1	3	.000	-.1393
Rater 4	10	2.00	.667	1	3	.000	.080

Table 4 shows descriptive statistics for four raters using a scale (1=easiest, 2=moderate, most difficult) to judge the agreement of the oral passages difficulty for each task (See Appendix A, section C).

In general, the majority of the tasks were rated moderate, indicating a reasonable difficulty of tasks. Descriptive statistics presented in this section represent the evidence for the content relevance which encompasses the different levels of difficulty/ ability for the oral passages of the SLE.

Construct

ANOVA analyses and planned comparisons were performed in order to find out if differences in students' performance at different proficiency levels were found in their SLE scores. Moreover, two one-way analyses of variance (ANOVA) procedures were performed to determine the extent to which the grammatical and sociolinguistic aspects were measured in the SLE. In all of these analyses, the dependent variables were the item difficulty estimates derived from the Rasch analyses. Analyses of contrasts between pairs of means were computed to test the three SLE construct hypotheses stated previously. All these analyses provided evidence to support the validity of the score interpretations about the language ability instructed in the Spanish courses which was specifically defined in our construct.

Table 5. Descriptive statistics for proficiency levels and IRT ability

		Mean in logits	SD	N
Novice-high	1	.93	.411	34
Intermediate-low	2	1.41	.522	59
Intermediate-mid	3	2.78	.593	51
Total		1.78	.927	144[ii]

Results for hypothesis about the proficiency level groups

As mentioned above, I hypothesized that test-takers with higher levels of Spanish proficiency would demonstrate higher performance on the SLE. In order to test this hypothesis, a one-way analysis of variance and planned comparisons was performed. In this analysis, students' a priori proficiency level according to the class they were enrolled in at the time of data collection was the independent variable and their IRT ability estimate on the SLE was the dependent variable. Table 5 shows the descriptive statistics for the IRT ability estimates of the three proficiency levels.

As can be seen in Table 5 the means indicate that students at the novice-high level are performing at a somewhat lower level compared to the other two proficiency levels, which was expected because this is the lowest proficiency level.

Table 6 shows that there was a highly significant main effect on the students' performance of three different proficiency levels. The F-radio for the linear unweighted (F=252.054, p=.000) indicates that as the proficiency level increased from 1 to 3 the ability also increased proportionally.

Table 6. Differences between IRT ability and three proficiency levels.

Ability	DF	MS	F	Sig.
Between Groups	2	41.964	151.614	.000
Linear Unweighted	1	70.041	252.054	.000
Within Groups	141	.277		

Table 7. Analysis of contrasts between proficiency levels and ability

Contrasts	t-value	df	t prob.
1/2	-4.240	141	.000
1/3	-15.908	141	.000
2/3	-13.646	141	.000

In order to better investigate the first hypothesis, analysis of contrasts examined three comparisons: one to test whether the basic proficiency level was different from the intermediate-low level, one to see whether the basic proficiency level was different from the intermediate-mid and one to see whether the intermediate-low was different from the intermediate-mid level (see Table 7).

Table 7 gives the statistics for each contrast. The three contrasts are used to test the hypothesis that novice-high (1) differs from intermediate-low (2) and intermediate-mid (3), and the intermediate-low differs from intermediate-mid. For all three contrasts we could say that there is an overall effect of proficiency level on ability.

These criterion-related analyses served as evidence from students' scores to support the assertion that the SLE measures the intended listening construct. What these criterion-related analyses mean is that the claims made on the basis of listening test results are supported by the test-takers' language proficiency.

Results for the hypothesis of linguistic characteristics

A one-way ANOVA with item difficulty as the dependent variable is used as evidence to show that the SLE measures grammatical knowledge. This analysis explores the item features to account for difficulty which is derived using the Rasch model. Difficulty is expressed in the measure of logits and is used as the dependent variable in the ANOVA. Five categories were identified by combining the type of items—local or comprehension—and the linguistic features of the items.

Table 8 shows the descriptive statistics for five categories of items. The means in logits show that the combinations of local and phonological items were the most difficult items for students. The easiest combination is the comprehension lexical items with the lowest mean in logits.

See below two examples for the comprehension item categories:
Participants listen to a monologue about a couple of friends visiting the library. Then, they read and are asked to respond to item 28 which was classified as a comprehension lexical item

Item 28: Comprehension lexicon

Seguramente a Blanca y a Juan les gustan los libros	*Cierto*	*Falso*
'Probably, Blanca and Juan like books'	True	False

Item 49: Comprehension syntax

Participants also listen to a description in which different buildings are compared and they have to deduce the following:

La farmacia tiene tanta luz como el supermercado	*Cierto*	*Falso*
'The pharmacy has as much light as the supermarket'	True	False

Table 8. Descriptive Statistics for five categories of item combinations

Categories	Mean	SD	N
Comprehension-lexical	-.8273	1.226	15
Local-lexical	-.1996	1.027	25
Local-syntactic	-.3694	1.161	16
Local-phonological	1.1671	.729	7
Comprehension-syntactic	.7942	1.059	19
Total	-.0006	1.242	82

Table 9. ANOVA results for five categories of item combinations

Difficulty	df	MS	F	Sig.
Between groups	4	8.742	7.480	.000
Within groups	77	1.1691		
Total	81			

Table 10. Results of contrasts between means in item combinations

Contrasts	t-value	df	t prob.
Comprehension/local	.967	77	.337
Lexical/phonological-syntactic	-4.012	77	.000
Lexical/Phonological	-4.063	77	.000
Lexical/Syntactic	-2.851	77	.006

A one-way ANOVA and an analysis of contrasts were performed in order to test the second hypothesis which states that comprehension items are more difficult than local items, and that the three different item categories are different from each other. Table 9 shows that there was a significant difference for all possible item combinations between the type and linguistic features. This finding is consistent with the SLE construct, since some item features are supposed to make items more difficult than others.

Table 10 shows the results of an analysis of contrasts between pairs of item combinations that were computed to examine whether there were significant differences between the pairs of item combinations: (1) comprehension items were compared to local items; (2)

lexical items were compared to phonological and syntactic items; (3) lexical items were compared to phonological items; (4) lexical items were compared to syntactic items.

Results of Table 10 indicated that no significant differences were found between the comprehension and local items, but there were significant differences between the lexical items and the other two features, between lexical and phonological items and between lexical and syntactic items at least at the 0.5 level.

Figure 1 shows that the mean difficulties for the local phonological combination are the highest and the most difficult, followed by the comprehension syntactic, local lexical, local syntactic and finally the comprehension lexical items represent the easiest combination.

Results for the third hypothesis based on sociolinguistic features

As the previous analyses of content relevance showed, oral passages were rated based on specific dialects spoken in the SLE tasks among other spoken features (see Appendix A section C). Regarding my third hypothesis, which stated that tasks classified as the most difficult based on the sociolinguistic features would demonstrate higher levels of difficulty on the SLE, a one-way ANOVA and analysis of contrasts were employed.

Ten tasks were rated from 1=easiest to 3=most difficult in terms of different features such as texts with idiomatic expressions or understanding dialects which were considered to be more difficult than texts spoken with standard Spanish. Table 11 shows the descriptive statistics of the difficulties of the tasks.

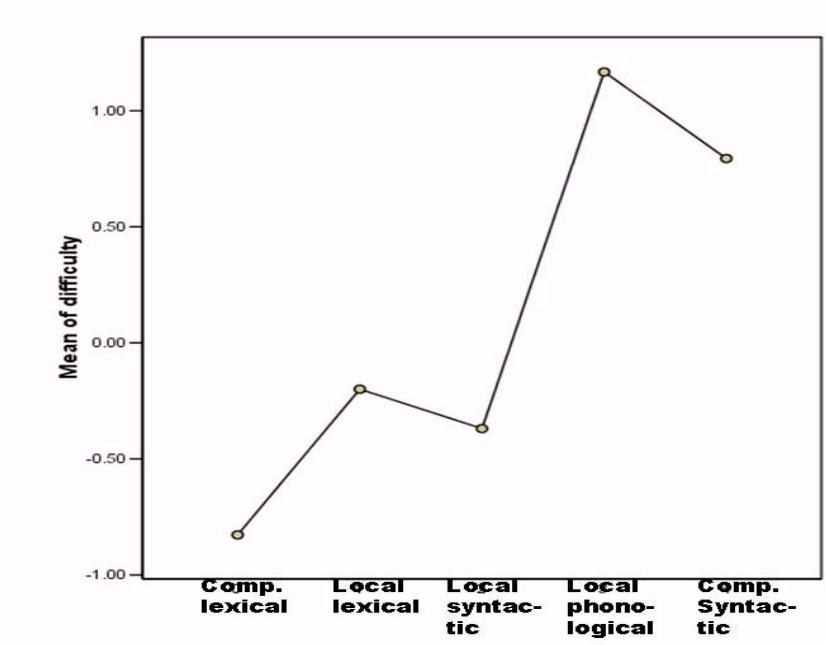


Figure 1. IRT mean of difficulty for all possible item combinations.

The mean difficulty of task level 1 and of level 2 are very similar but the mean difficulty for level 1 is lower than that for level 2. The mean difficulty for level 3 is higher than that for levels 1 and 2. Table 12 shows a significant difference among the level of tasks as predicted during the test development (F= 4.368, df = 2, p=.016).

In order to know more about these effects between tasks, the following planned comparisons were employed: one to test whether the easiest tasks were different when compared to the moderate tasks (1/2), one to see whether the easiest tasks were different when compared to the most difficult task (1/3) and one to see whether the moderate tasks were different when compared to the most difficult task level (2/3). Results appearing in Table 13 indicated that the easiest tasks were not significantly different from the moderate tasks, but they were significantly different from the most difficult tasks. At the same time, moderate tasks were significantly different from the most difficult tasks. So, we could say that comparing task at the level 1 and 2 the difficulty level of these comparisons is the same. There is no clear cut difference between those levels because the difficulty levels overlap. However, there were significant differences when comparing task level 1 to 3 and level 2 compared to 3.

Table 11. Descriptive Statistics for task level

Task level	Mean	SD	N
1	-.3341	1.40924	17
2	-.1382	1.15638	49
3	.7750	1.05050	16
Total	-.0006	1.24205	82

Table 12. Results of ANOVA for task levels and difficulty measure

Difficulty	df	MS	F	Sig.
Between Groups	2	6.222	4.368	.016
Within Groups	79	1.424		
Total	81			

Table 13. Contrasts between means in task levels

Contrasts	t-value	df	t prob.
1/2	-.583	79	.561
1/3	-2.668	79	.009
2/3	-2.657	79	.010

CONCLUSION

Multiple types of evidence were gathered to support the construct validity of interpretations made from the SLE: content relevance, criterion-relatedness, and content coverage. The evidence of criterion-relatedness and content coverage was gathered by performing an ANOVA and *a priori* comparisons to determine whether construct-related hypotheses based on previous studies of listening comprehension pointed at the expected patterns. Evidence of content relevance was established by verifying that the SLE tasks actually consist of different linguistic points that were appropriate for different listening levels.

Descriptive statistics for the four raters on the oral passages derived from materials in Spanish t**e**xtbooks, the ACTFL scale, and other sociolinguistic attributes, which represented the evidence for content relevance, indicated that the raters were in agreement on the three scales with little variation in their ratings of the listening passages. This suggests that the evidence gathered via the raters' evaluation is consistent based on three different rating scales.

Evidence was found to support all three hypotheses about test performance. First, expected differences were found between the three proficiency level groups. This confirms the first hypothesis that students in the higher proficiency levels would obtain the higher test scores on the SLE. This finding of criterion-relatedness supported the intended interpretation about test-takers' listening ability and the intended use of placing students into different Spanish courses based on their Spanish proficiency level.

The second hypothesis was that comprehension items would be more difficult than local items for test-takers. In addition we had posited that lexical features would be different from phonological or syntactic features. Findings showed that our second hypothesis was partially consistent with results from Shohamy and Inbar's (1991) study. They found that comprehension items, which they called global items, were more difficult than local ones: however, our findings revealed no significant differences between comprehension and local ones. One explanation for the lack of consistency with Shohamy and Inbar's study might be that the characteristics of our construct included easy items. This decision was taken during test development in order to insure that beginners would not be discouraged by not understanding difficult oral passages. Thus, 10 out of 15 of the comprehension items presented a lexicon that was easy for beginners with well-known linguistic items such as 'art', 'attendance', 'variation', 'to like', 'family', 'good' and other words that are taught in the first year. Not surprisingly, the test-takers were able to process the lexical input of the items and were more likely to get the correct answer. This rationale could have caused the finding of non significant differences between local and comprehension items. Moreover, the format for 14 comprehension lexical items was true or false selection rather than limited response where students need to type a Spanish word. The format of the items is also an important variable in order to find differences. Comprehension combined with lexical features was demonstrably easier due to the lexical nature of the items, but comprehension of syntactic items constituted the second

most difficult items. The nature of the syntactic items was most probably accounted for by the fact that the beginners did not get the correct answer since 12 out of 19 comprehension syntactic items were intended for intermediate and advanced learners. This finding suggests that linguistic features can also be used as a discriminator of L2 listening performance.

Part of our second hypothesis (i.e., comprehension items would be more difficult than local items…) was confirmed when significant differences between the linguistic items were found, in keeping with Shohamy and Inbar (1991) and Tsui and Fullilove's (1998) results on L2 listening. Shohamy and Inbar (1991) found that test-takers with less listening ability tend to understand the passages by interpreting local items. That is, learners tend to focus on the linguistic input to get the correct answer. Tsui and Fullilove (1998) found that linguistic input is very important in order to discriminate among the L2 listeners. They found out that poor listeners tended to guess or focus on their background knowledge when they did not have the necessary linguistic knowledge to answer questions correctly. Despite the fact that all of our items were coded with a combination of comprehension or local items and a linguistic feature, items with lexical, phonological or syntactic features were compared to each other regardless of their local or comprehension classification. Significant differences were found in the Rasch difficulty estimates among lexical, phonological and syntactic items with the following pattern of difficulty: lexical < syntactic < phonological.

The third hypothesis was that items classified as the most difficult tasks based on the sociolinguistic features would yield higher mean difficulty estimates than those in the easiest tasks. Results from a one-way ANOVA and analysis of contrasts confirmed our third hypothesis. A significant difference in the sociolinguistic features of the listening tasks was found. In addition, analysis of contrasts between pairs of tasks indicated that there were significant differences between the easiest tasks and the most difficult tasks, as well as between moderate tasks and the most difficult tasks. However, no significant differences were found between the easiest tasks and moderate tasks. This finding suggests that sociolinguistic variables play an important role in distinguishing between beginners and advanced listeners.

The construct validity evidence reported in this paper was an important part of the assessment use argument that includes other types of evidence as well. An assessment use argument (Bachman 2005) is composed of two parts: 1) the validity argument and 2) the utilization argument. The construct validity approach presented in this paper is just one component of the six mentioned in the validity argument. For more information about the context of the research and the complete assessment use argument see unpublished dissertation of Pardo Ballester (2007).

REFERENCES

ACTFL. (1986). Retrieved on May 2003 from http://www.sil.org/lingualinks/languagelearning/OtherResources/ACTFLProficiencyGuidelines/contents.htm

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*, 1-34.

Blake, R. J., González Pagani, M. V., Ramos, A., & Mraks, M. A. (2003). *Al corriente* (4th ed.). New York: McGraw-Hill.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, 8, (1), 67-91.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Davidson, F. & Lynch, B. (2002). *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven: Yale University Press.

Linacre, J. M. (2006). Winsteps (Version 3.61.2) [Computer Software]. Chicago: Winsteps.com.

Pardo Ballester. (2007). The Development of a Web-Based Spanish Listening Placement Exam. Unpublished PhD Dissertation. University of California, Davis.

Shohamy, E. & Inbar O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language testing, 8*, 23-40.

SPSS (2003). SPSS version 12 for Windows. Chicago, IL: SPSS Inc.

Terrell T., Andrade, M., Egasse, J. & Muñoz E. (1998). *Dos Mundos* (4rd ed.). New York: McGraw-Hill.

Tsui, A. B. M., and Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics, 19*(4), 432-451.

UCLA, Department of Applied Linguistics and TESL & Center for Digital Humanities. (2003). WebLAS (Web-based Language Assessment System). Retrieved July 17, 2006, from http://www.weblas.ucla.edu/.

APPENDIX A

TEXT CONTENT EVALUATION

SPANISH LISTENING EXAM
EVALUATION FORM FOR 10 PASSAGES

For each of the passages below, circle the number that REFLECTS YOUR VIEWPOINT on a five-point scale where:

1=Strongly disagree
2=Disagree
3=Undecided
4=Agree
5=Strongly agree

SECTION A: TEXT CONTENT

The topics of the SLE passages are drawn from what Spanish instructors might do in a typical class for instructional purposes. Topics and grammar are adequate for 1st Spanish year (most of them are drawn from Dos Mundos).

Task 1: Culture taught in Spanish 2 (Lesson 7) 1 2 3 4 5

Task 2: The content could be heard during the first week of Spanish 1, 2, or 3 with some modifications.
1 2 3 4 5

Task 3: This topic is studied in Spanish 3 (lesson 13). The content of this passage presents grammar learned in Spanish 2 (conditional) and Spanish 3 (Subjunctive)
1 2 3 4 5

Task 4: Descriptions of people are learned in Spanish 1, 2, and 3
1 2 3 4 5

Task 5: This topic and content could be presented at any time during the quarter in Spanish 1, 2, 3, 21, 22, and 23. The dialect presented in this task is from Spain, using the 'vosotros' form. Any Spanish instructor from Spain or one that is used to this dialect could inform their students (if needed) with similar information. Also any Spanish instructor, no matter their dialect could use this topic if necessary.
1 2 3 4 5

Task 6:Spanish 2 (topic: Lesson 9) 1 2 3 4 5

Task 7: Spanish 2 (Grammar and topic-comparisons-learned in lesson 6)

1 2 3 4 5

Task 8: Spanish 2 (topic and grammar learned in lesson 9) Grammar: imperfect tense.

1 2 3 4 5

Task 9: Spanish 3 (topic and grammar learned in lesson 11) conditional tense, present indicative and subjunctive

1 2 3 4 5

Task 10: Spanish 1 (culture learned in lesson 4) Grammar: present tense indicative and subjunctive, present perfect indicative and conditional.

1 2 3 4 5

SECTION B: TEXT CONTENT

The Spanish department at UC Davis classified the following Spanish classes at the following ACTFL level:

Spanish 2: Novice-high
Spanish 3, 21: Intermediate-low
Spanish 22, 23: Intermediate-mid

Novice-High (Spa 2)

Able to understand short, learned utterances and some sentence-length utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends words and phrases from simple questions, statements, high-frequency commands, and courtesy formulae. May require repetition, rephrasing, and/or a slowed rate of speech for comprehension.

Intermediate-Low (Spa 3 y 21)

Able to understand sentence-length utterances which consist of recombinations of learned elements in a limited number of content areas, particularly if strongly supported by the situational context. Content refers to basic personal background and needs, social conventions and routine tasks, such as getting meals and receiving simple instructions and directions. Listening tasks pertain primarily to spontaneous face-to-face conversations. Understanding is often uneven; repetition and rewording may be necessary. Misunderstandings in both main ideas and details arise frequently.

Intermediate-Mid (Spa 22 & 23)

Able to understand sentence-length utterances which consist of recombinations of

learned utterances on a variety of topics. Content continues to refer primarily to basic personal background and needs, social conventions and somewhat more complex tasks, such as lodging, transportation, and shopping. Additional content areas include some personal interests and activities, and a greater diversity of instructions and directions. Listening tasks not only pertain to spontaneous face-to-face conversations but also to short routine telephone conversations and some deliberate speech, such as simple announcements and reports over the media. Understanding continues to be uneven.

Please rank the passages with 1 for Novice-high level, 2 for intermediate-low level and 3 for intermediate-mid level for measuring listening comprehension

Passages	Rank
1.	_____
2.	_____
3.	_____
4.	_____
5.	_____
6.	_____
7.	_____
8.	_____
9.	_____
10.	_____

SECTION C: TEXT CONTENT

The following listening texts were ranked according to the rate of delivery with:

1 for beginners (a maximum of 120 words per minute)
2 for intermediate learners (a maximum of 160 words per minute)
3 for advanced learners (200 words per minute)

Please rate again the texts (with 1 being the easiest, 2 moderate and 3 the most difficult) considering the following:

1) your viewpoint ranking these texts according to the ACTFL guidelines
2) the rank of the rate of delivery
3) texts more familiar to the listener tend to be easier
4) texts with fewer things or people to be distinguished tend to be easier
5) texts with concrete content tend to be easier
6) texts with idiomatic expressions or understanding dialects tend to be more difficult than texts spoken with standard Spanish

Texts	Rank Rate of delivery	New Rank
1.	___2____	_____
2.	___2____	_____
3.	___2____	_____
4.	___1____	_____
5.	___2____	_____
6.	___1____	_____
7.	___2____	_____
8.	___3____	_____
9.	___3____	_____
10.	___3____	_____

[i] For security of the SLE, the items presented here are not completely the same as the ones used in the SLE.
[ii] Note that three participants were dropped because of cheating problems.

About the Authors

Cristina Pardo-Ballester (Ph.D., University of California, Davis) is Assistant Professor of Spanish in the World Languages and Cultures Department at Iowa State University. Her primary areas of research are second language acquisition (SLA), Computer-Assisted Language Learning (CALL), and Language Assessment. She is currently working on the development of Spanish hybrid courses integrating different uses of technology in Spanish instruction. One of her goals is to create better listening tests by including the use of visual, collaborative listening and computer-based testing. The integration of the hybrid courses into the Spanish curriculum program at ISU is intended to assess students' performance, students' attitude and students' voice about learning with environments, the technology and the contact with humans.

Doe-Hyung Kim (MA, University of Illinois at Urbana-Champaign) is Project Coordinator for the Curriculum, Technology, and Education Reform (CTER) online Master of Education program at the University of Illinois. His research interests include feedback and behavior tracking in computer-assisted language learning, electronic corpora for ESL writing, and online learning. He maintains the CTER online program, which in its 10th year, serves practicing teachers nationwide wishing to integrate technology in classrooms by offering graduate courses online. He is currently working on a Flash/Actionscript-based program for testing explicitness of feedback for improving ESL grammar.

Elena Cotos is a doctoral student in the Applied Linguistics and Technology program at Iowa State Univeristy. Her research interests include automated scoring, learner corpora, computer assisted language learning, computer assisted language testing, academic writing, and materials design. Elena has presented at CALICO, TESOL, and TSLL, and has reviewed for the TESOL Quarterly. She has also designed online materials for the English Listening Lounge and participated in the development and implementation of "Life in a Second Language" Simulation and Enhancing and Advancing Science for English Language Learners projects.

Eunice Eunhee Jang (Ph.D., University of Illinois at Urbana-Champaign) is Assistant Professor at the Ontario Institute for Studies in Education at the University of Toronto. Her research interests include cognitive diagnostic assessment, validity and fairness of educational and language assessment, Differential Item Functioning, and test dimensionality. She is currently working on integrating formative diagnostic assessment into ESL literacy instruction. She also collaborates with Dr. Jim Cummins on the validity project for "Steps Toward English Proficiency" with over 50 ESL teachers across Ontario. STEP is intended to serve teachers to assess and track all English language learners' literacy development in Ontario schools.

Jinhee Choo is currently a graduate student at the department of Educational Psychology at University of Illinois at Urbana-Champaign. Her field of experience includes second language acquisition, language data analysis, corpus linguistics, teaching ESL and computer-assisted language learning. She participated in a CALL project, the *ESL Tutor* to aim at eliminating typical Korean ESL learners' errors from written compositions.

John M. Levis (Ph.D., University of Illinois at Urbana-Champaign) is Associate Professor of TESL/Applied Linguistics at Iowa State University. His research interests include speech intelligibility, intonation, and English pronunciation. His articles have appeared in *TESOL Quarterly, World Englishes, Applied Linguistics, ELT Journal, System, TESOL Journal, Language Awareness,* and *Annual Review of Applied Linguistics*. He is currently writing a book on English pronunciation.

Maja Grgurovic is a doctoral student in the Applied Linguistics and Technology Program at Iowa State University. She holds an MA degree in TESL/Applied Linguistics from Iowa State. Her research interests are CALL, multimedia, materials design, and integration of technology into language teaching and teacher education. Maja has presented at TSLL, CALICO, WorldCALL, TESOL, and SLRF and published in Language Learning and Technology, ReCALL, and TESL-EJ.

Mathias (Mat) Schulze is Associate Professor of German at the University of Waterloo in Ontario. He obtained his PhD in Language Engineering at UMIST in Manchester (England), co-authored a book on the application of artificial intelligence to computer-assisted language learning (Heift and Schulze, 2007), and has published a number of papers on Computer-Assisted Language Learning and grammar. His main interests are in the application of artificial intelligence techniques, such as student modeling and natural language processing, and second language acquisition research to CALL.

Melissa Baralt is a PhD candidate studying Second Language Acquisition in the Spanish and Portuguese Department at Georgetown University. Her research interests include technology incorporation in SLA, bilingualism, and task-based language teaching.

Nathan T. Carr (Ph.D., University of California, Los Angeles) is an Assistant Professor at California State University, Fullerton in the TESOL Program, part of the Department of Modern Languages and Literatures. His primary research interests include computer-based testing, particularly with respect to automated scoring of constructed response tasks; writing materials for training teachers in language testing; and validation studies, particularly those involving the relationship between test task characteristics and examinee performance. He is also involved in his department’s efforts to develop and validate proficiency tests in a variety of languages.

Nick Pendar (Ph.D., University of Toronto) is Assistant Professor with the Applied Linguistics & Technology and Human Computer Interaction programs at Iowa State University. His specialty is computational linguistics and natural language processing.

His research interests include the use of machine learning techniques in natural language processing, as well as intelligent computer assisted language learning and automated scoring.

Quan Zhang (Ph.D., Guangdong University of Foreign Studies at Guangzhou) is Chair Professor at College of Foreign Studies, Southern Med. Univ. in Guangzhou, China. His research interests include cognition and testing, computerized cognitive assessment, IRT computer software and linguistics. He is currently working on SEM and EQS for language testing and moderating his Cognitive Response Theory. In 2002, he was invited as senior visiting scholar to ETS. From July 2006 to July, 2008, he is invited as a research scholar to Department of Applied Linguistics &TESL, UCLA. In China, he is a good collaborator with Prof. Lyle F. Bachman.

Robert Mislevy (Ph.D., University of Chicago) is Professor of Measurement, Statistics, and Evaluation (EDMS) at the University of Maryland. He applies developments in statistical methodology and cognitive research to practical problems in educational and psychological measurement. His work has been recognized with honors and awards such as the American Educational Research Association's Raymond B. Cattell Early Career Award for Programmatic Research, the National Council of Measurement in Education's Award for Technical Contributions to Educational Measurement (three times), the ETS Senior Research Scientist Award, and the International Language Testing Association's Samuel J. Messick Memorial Lecture Award. In 1992, he was elected president of the Psychometric Society and nominated as a Fellow of the American Psychological Association, and in 2003 he was presented the National Council of Measurement's Award for Career Contributions to Educational Measurement. His work has included a multiple-imputation approach for integrating sampling and test-theoretic models in the National Assessment of Educational Progress (NAEP), a Bayesian inference network for updating the student model in an intelligent tutoring system, and a demonstration of a framework for monitoring and improving portfolio assessment evaluation (in the context of the Advanced Placement Studio Art Portfolio assessment).

Tony Becker is a Ph.D. student at Northern Arizona University in the Applied Linguistics Program. His research interests include assessment and computer-assisted language learning. His dissertation interest focuses on the assessment of declarative and procedural knowledge in online computer tests.

Viviana Cortes (Ph.D., Northern Arizona University) is Assistant Professor in the TESOL/Applied Linguistics Program in the English Department at Iowa State University in Ames, Iowa. Her research Interests include corpus-based studies of register variation, and the study of the use of fixed word combinations in different academic registers. Her latest articles can be found in English for Specific Purposes, Applied Linguistics, Linguistics and Education, and the Journal of English for Academic Purposes, as well as in various edited volumes.

Xiaoming Xi is a research scientist in the Research & Development Division at Educational Testing Service. She earned a doctorate degree in second/foreign language assessment from the University of California, Los Angeles. Her areas of interest include factors affecting performance on speaking tests, rating scales for speaking tests, rater bias issues in speech scoring, automated scoring of speech, and validity and fairness issues in the broader context of test use. Xiaoming has published in leading journals and wrote a chapter on "Methods of Test Validation" for the second edition of the Encyclopedia of Language and Education. She also serves on the Editorial Boards of Language Testing and Language Assessment Quarterly. She is a recipient of the 2002 Lado Best Student Paper Award, the 2003 Spann Fellowship Award for Second/Foreign Language Assessment, the 2005 ILTA Best Paper Award, and the 2005 and 2006 ETS Presidential Award.